DOCUMENT RESUME

ED 213 733 TM 820 078

AUTHOR Linn, Robert
TITLE Conceptualization of Issues in Construct and Content Validity. Studies in Measurement and Methodology, Work Unit No. 1: Conceptual and Design Problems in Competency-Based Measurements.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Jun 78
NOTE 96p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Item Analysis; Item Sampling; Test Bias; *Test Construction; *Testing Problems; Test Items; *Test Reliability; *Test Validity

ABSTRACT

A series of studies on conceptual and design problems in competency-based measurements are explained. The concept of validity within the context of criterion-referenced measurement is reviewed. The authors believe validation should be viewed as a process rather than an end product. It is the process of marshalling evidence to support interpretations. Performance standards are important in the applications of criterion-referenced measurement. Shepard suggested standard setting should be an iterative process and one that involves various audiences. The problem of deciding on test length and passing score required to make sufficiently accurate mastery decisions is discussed. Self scoring, or computer administered, tests might provide savings in average testing time by making decisions sequentially. Item analyses for criterion referenced measures are likely to be as relevant to investigations of instructional processes as they are to concerns about item characteristics. In principle, item analysis results should not be used for item selection. Reliability procedures based upon assessment of consistency of decisions appeal to the authors as possessing the greatest utility. Estimation of domain scores, test bias, and matrix sampling are reviewed at length. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED213733

DELIVERABLE: June 1978

STUDIES IN MEASUREMENT AND METHODOLOGY

WORK UNIT 1

Conceptual and Design Problems in Competency-Based Measurements

Eva Baker
Project Director

CONCEPTUALIZATION OF ISSUES IN CONSTRUCT AND CONTENT VALIDITY

Robert Linn

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. S. Gray
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 078

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California - Los Angeles

June 1978

CONCEPTUALIZATION OF ISSUES IN CONSTRUCT AND CONTENT VALIDITY

Robert Linn
University of Illinois at Urbana-Champaign*

*Work performed while at CSE as a Visiting Scholar

"The activity which is the subject of this report was supported in part or in whole by the National Institute of Education, Department of HEW. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred."

The concept of validity has long been one of the most fundamental concepts in educational and psychological measurement. Much has been written about the various types of validity and the process of validation. Yet, within the context of criterion-referenced measurement, validity has received relatively little attention. Certainly it has not received the amount of attention that has been given to various other topics such as reliability, determination of test length, and questions of the need for variability.

The relative lack of attention to issues of validity for criterion-referenced tests may be largely attributed to a view that of the three types of validity discussed in the Standards for Educational and Psychological Measurement (APA, 1974) only concent validity is particularly relevant to criterion-referenced measures. The key steps in content validation are first the definition of domain and, second, the demonstration that the measure adequately represents the domain (Linn, 1977). If the preceding is correct, then it is hardly surprising that validity has not been the topic of hot debate in a criterion-referenced context that it has been in other measurement contexts.

The great emphasis that has been placed on domain definition does much to solve the problem of content validation. Where the domain is sufficiently well defined, random sampling of items can provide a means of achieving representativeness. This is certainly a much stronger basis for making inferences about a content domain than is provided by traditional test construction procedures. Thus, issues of validity may have seemed less salient for criterion-referenced measures because, when based on adequate definition and sampling, the type of validity of primary concern

is assured.

Of course, the ideals of a completely defined domain and construction of measures by random sampling are rarely even approximated in practice. Even content validity will often be problematic. Furthermore, it is our contention that, like other measures, criterion-referenced measures are subject to many interpretations and uses. The various interpretations and uses require the consideration of a much wider range of evidence than would ordinarily be considered under the heading of content validity. Thus, we shall argue that validity is a topic worthy of much more consideration than it has received in most discussions of criterion-referenced measurement.

Types of Validity

The validity of a test is often defined as the degree to which it measures what it is supposed to measure. A broader and more useful conception, however, is that questions of validity are questions of the soundness of interpretations of a measure (Cronbach, 1971; APA, 1974). Validation is best conceived of as a process rather than an end product. It is the process of marshalling evidence to support interpretations. Measurement results can have many interpretations which differ in their validity and in the type of evidence required for the validation process.

There are a variety of types of questions of validity which require different types of validation, e.g. criterion-related, content and construct. It is convenient to distinguish different types of validation. Certainly, many labels and category systems have been offered by various authors who have written about validity. It is worth noting, however, that "these aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation" (APA, 1974).

The Variety of Validation Needs for Criterion-Referenced Measures

Millman (1974) described validity in terms of the "...accuracy of inferences made from test scores" (p. 360). This position is generally consonent with the one stated above. While recognizing a variety of types of inferences requiring different validation processes, Millman argued that the most appropriate inference for a domain-referenced test concerns the status of an individual relative to a well defined domain. For this inference, the logical analysis of the domain definition and procedures for generating items are of primary concern. That is, content validation is the primary issue. Hambleton, Swaminathan, Algina and Coulson (1975) stated a similar position. They acknowledged that other types of validity are appropriate to study, but claimed that content validity is the "center of validation concerns" for a domain-referenced [here, criterion-referenced] test.

We do not quarrel with the emphasis placed on content validity by Millman and by Hambleton, et al. Indeed, we think that one of the main advantages offered by criterion-referenced measurement is that it provides a much firmer basis for content validation than has been available with traditional psychometric approaches to constructing achievement tests.

Content validity is a necessary consideration for criterion-referenced measures. It is our contention, however, that it is rarely sufficient. This is so because rarely, if ever, are inferences restricted to the type that are adequately supported by content validity alone.

With a well defined domain and random sampling of items from that domain there is a firm basis for supporting an inference that a person can satisfactorily perform at least x percent of the tasks in that domain. This is the type of inference that is supported by content validity. It is a natural consequence of the adequacy of the domain definition and of item

generation rules. Seldom are our inferences limited to such a degree, however. Labels such as "master" or "competent" are often attached to persons whose performance is estimated to be above some specified level. Scores on the test are apt to be used as the basis of differential treatment. For example, some students may be given remedial instruction as a result of their scores. These additional interpretations and uses require more than content validity (Linn, 1976).

When a person is labeled a "master" certain expectations usually accompany that label. The student who has mastered certain types of addition problems may be expected to be "ready" to learn subtraction. Mastery of one skill may be expected to be a prerequisite for mastering another skill. Evidence of the soundness of expectations such as these is needed. The process of marshalling the evidence for the soundness of the expectations is construct validation (Cronbach, 1971; Messick, 1975).

Even where constructs are not invoked more than content validity will usually be needed. The decision to have students with scores on a test below some specified level repeat an instructional segment rather than go on to a new one is based on an implicit prediction that the student will, at least, learn the content of the segment to be repeated better by redoing it than by moving on. It may also be based on an implicit prediction that he or she would do more poorly on the next segment of instruction if allowed to move on to it before repeating the earlier segment than would otherwise be the case. Evidence of criterion-related validity is needed to support these predictions.

We have argued that all three of the traditionally recognized types of validity are relevant to criterion-referenced measurement just as they are to other types of measurement. It does not follow from this position, however, that the traditional validation procedures will be the most appropriate

CSE
ARCHIVE COPY

ones for criterion-referenced measures. For instance, there are a variety of reasons that the traditional reliance on correlation coefficients may be inappropriate for criterion-referenced measures. One of these is the issue of variability which is discussed at length in the reliability section of this paper. Another reason is that, as suggested by Hambleton and Novick (1973), correlations are based on an inappropriate loss function for situations commonly encountered with criterion-referenced measures. Hambleton and Novick suggested that a threshold loss is more appropriate in a criterion-referenced context than is the traditional squared-error loss.

The need for a different type of statistics or for nontraditional procedures for collecting the needed data do not require different concepts of validation. They merely represent different details of implimentation within a common conceptual framework. A few comments regarding the special implimentation requirements for validation of criterion-referenced measurement may be useful, however. These are offered below under the three traditional headings of content, criterion-related, and construct validity.

Content Validity

Domain definition

Although content validity is formally recognized as one of the three main types of validity (APA, 1974), few would consider it to stand on an equal footing with the other two types of validity in terms of the rigor of the evidence that is usually provided to support a claim of validity. Indeed, some well known test theorists have argued that what traditionally goes under the heading of content validity shouldn't even be called validity. For example, Messick (1975) suggested that

we should "call it 'content relevance', ... or 'content representativeness', but don't call it content validity" (pp. 940-941). Ebel (1975a) suggested that "perhaps... we should call it content reliability ..." (as quoted by Guion, 1977, p. 2).

Guion (1977) clearly stated a number of reasons for having reservations about content validity. Primary among these reasons is Guion's conclusion that "judgements of content validity have been too swiftly, glibly and easily reached in accepting tests that otherwise would never be deemed acceptable" (1977, p. 8). Despite his reservations, Guion argued that the ideas contained under the notion of content validity are extremely important.

For the acceptance of a measure on the basis of content validity Guion proposed a set of five minimal conditions. These conditions are:

1. that the content domain must involve "behavior with a generally accepted meaning" (p. 6),
2. that the definition of the domain must be unambiguous,
3. that the domain must be relevant to the purposes of the measurement,
4. that "qualified judges must agree that the domain has been adequately sampled" (p. 7), and
5. that the measure must have reliability.

The above list is useful. But, it involves considerations that go beyond what we would include under the heading of content validity. We will argue that content validity is derived from only two considerations: domain definition and representativeness. Our disagreement with Guion may be more one of semantics than substance however. Issues on "meaning" and "relevance" imply inferences that go beyond those justi-

fied on the basis of content validity alone. They involve constructs or external criteria and require other forms of validity evidence. As we have already argued, content validity is necessary but not sufficient to support "the acceptance of a measure" for any but the most limited types of used and inferences.

The key to content validation is the definition of the domain of interest. Ideally, the definition should include a description of every detail of the measurement procedure that might have an influence on the results. The ideal can never be fully achieved in practice, but improvements over common practice could certainly be made.

Consider, for example, the observations of student teachers in a classroom setting. There is an almost limitless list of potentially important details that might have an influence on the results of the observation procedures. There are the physical conditions of the situation(e.g., building, lighting, room size, furniture arrangement, etc.) There are considerations

of timing (e.g., time of day, day of week, part of the semester). There are the students and other persons in the room to be considered (e.g., number, age, sex, etc.). The topic of the lesson, the amount of previous contact of the student teacher with the students, the obtrusiveness of the observers, and the characteristics of the observers all require attention. The list could go on and on. The description can never be complete. Judgment is always required in determining what is most deserving of attention, i.e., those aspects of the measurement procedure that are most crucial to the results of the measurement.

Certain aspects of the procedure will be controlled. For example, in the observation of student teachers, the content of the lesson, the number of students and the length of time available for the lesson, among a number of other conditions, might be fixed for everyone. All the possible variations of other conditions not fixed by the definition, e.g., observers, physical location, etc, constitute what Cronbach, et al. (1972) refer to as "the universe of admissible observations."

It is probably more common in discussions of content validity to speak in terms of a "performance domain" (e.g., APA, 1974) rather than a universe of admissible observations. The latter label, though longer and more clumsy to use all the time, has the advantage of calling attention to the idea that any characteristic of the measurement procedure which is not controlled in the definition, could be varied and the observation would still be admissible. For example, the domain of all addition problems involving two single digit positive numbers appears fairly clear and specific. Such a definition, however, would include as admissible observations oral as well as written presentation, base eight as well as base ten numbers, speeded as well as unspeeded presentations, multiple-choice as well as constructed response formats, and many other variations. The universe of admissible

observations calls attention to the limits and possible vagueness of the definition of the measurement procedure that may go unattended in the description of a performance domain.

Domain Representativeness

As noted above, definition of the universe of admissible observations is the first of two crucial steps in content validation. The second step requires a demonstration that the observations yield results that are representative of that universe. Where the universe is sufficiently well defined, representativeness may be pursued by a formal sampling process. Random or stratified random sampling has great advantages of objectivity and enables one to calculate unbiased estimates of population parameters. For example, if test items are randomly sampled from a universe of items, the proportion of items correct in the sample of items administered to an individual is an unbiased, maximum likelihood estimate of the number of items in the universe that the individual could answer correctly (Harris, Pearlman & Wilcox, in press). Furthermore, it does not depend on who else was tested, either as an estimate or for interpretation.

Defining universes such that random samples of items or other types of observations can be selected has great advantages but is seldom accomplished. Some have argued that it doesn't even approximate reality and that it is misleading to use a sampling model (e.g., Loevinger, 1965). At least for some limited examples, however, domains have been defined such that sampling is possible. The work of Bormuth (1968, 1970) and of Hively and his associates (e.g., Hively, Patterson & Page, 1968) provide some important examples. Some of the potential advantages of this type of approach in the area of achievement testing have been discussed by Shoemaker (1975).

Lip service is often given to the requirements of clear definition. It is fundamental to any approach that is concerned with "representative" or

accepted as desirable (see for example p. 28 of the APA Standards for Educational and Psychological Tests, 1974). Unfortunately, satisfactory definitions are the exception rather than the rule.

There are a lot of reasons that our measures lack the sort of rigorous definitions and the sampling processes that are required. For many content areas, the problem is a very difficult one, certainly more difficult than the area of simple arithmetic problems which has become the overused example. It is also the case that test publishers have relied too heavily on a totally different approach to achieving meaning. Norms rather than content definition are used to provide meaning. Note that it is not, as some have implied, the existance of norms as a basis of comparison that is the problem. Any test can be normed. Rather it is the reliance on norms as the sole basis of interpretation that is the culprit.

The need for content interpretations of achievement tests has been recognized for a long time. Ebel (1962) argued forcefully for content interpretations and provided an example where the universe and sampling procedures were defined with such rigor that it made no difference whether he or his secretary constructed the test. Ebel concluded that "a test produced by objectively defined processes may be less efficient, or lack some kinds of excellence which a creatively artistic test construction might achieve, but the increase in objective meaningfulness and reproducibility could more than offset the cost" (1962, p. 22).

Applying the principles of universe definition and sampling to measures such as observations of teacher behavior, or ratings of student art projects is substantially more complicated than applications with spelling or arithmetic tests. The necessity of doing so, however, is just as great if we are to achieve the sort of "objective meaningfulness and reproducibility" described by Ebel. For measures, such as teacher observations, the universe of admissible

such a measure could be assessed by duplicate-construction experiment. Two independent teams would use the same definition of the universe of admissible observations to construct the measures. The results of applications of the two independently constructed measures would then be compared. The ideal outcome would be that the results of the two measures were always equivalent except for sampling errors. The content validity of the measure would then be assessed in terms of the degree to which this ideal is approximated.

Why Call It Validity?

After describing what we consider to be the two essential aspects of content validity we must return to a nagging concern raised in the beginning of this section. What we have argued for might, as Ebel (1975a) suggested, better be called "content reliability". There is a close correspondence between the notions of generalizability theory and those of content validity as we have espoused them. We have retained the label of content validity for two reasons.

First, the conotation of reliability is generally narrower than what we think should be included under considerations of content validity. As commonly thought of, reliability need not involve the careful attention to unambiguous definition and sample representativeness. The common procedures for estimating various types of reliability, (e.g. internal consistency, parallel forms reliability, or test-retest reliability) can be applied with little concern for the clear definition of the limits of the domain or for questions of representativeness.

Our second reason for preferring to retain the content validity label is that it makes the tie among the three types of validity more apparent. Content validity can seldom stand alone. On the other hand, it is a

observations may be defined in terms of several independent characteristics, which Cronbach, et al. (1972) refer to as facets. For example, one facet might be observers. A population of potential observers would need to be defined from which random samples could be selected. Other facets might be occasions, subject matter of lesson, and items on which observers make observations. In practice, it may seldom be possible to actually draw random samples of observations representing each facet. For example, location and scheduling may make some compromise necessary in the assignment of observers. Does this mean that the notions of universe definition and sampling have no place in the discussion? We think not. As noted by Cronbach:

> "...it is customary to apply statistics derived from a random sampling model to groups of persons who were not truly drawn at random and are only loosely representative of the population of persons to which the investigation generalizes. Sometimes the mismatch between model and reality is serious, sometimes not, but the problem is not peculiar to content sampling" (1971, p. 455).

Cornfield and Tukey's (1956) famous analogy between a bridge with two spans and statistical and substantive aspects of data interpretation is apropos here. The statistical span of the bridge goes from one shore to an island and the substantive span from the island to the other shore. They argue that where the substantive span is weak, it may be better to have that span of the bridge short while stretching the statistical span.

The absence of the ability to randomly sample from a domain does not preclude the possibility of studying generalizability. Cronbach (1971) suggested a procedure that would provide a check of the content validity of a measure where components of the measure were selected judgmentally rather than by a set of sampling rules. He suggested that the content validity of

necessary accompaniment to the other types of validity. For example, content validity of a test may be given little consideration in a traditional criterion-related validity study but may the central basis for the acceptance of the criterion measure. To do otherwise could lead to the need for a criterion against which the first criterion ~~against which the first criterion~~ could be validated and the continued application of that logic only leads to an infinite regress (Guion, 1977, p. 8).

Criterion-Related Validity

Criterion-referenced tests are often used to make short term instructional decisions. Commonly a dichotomy is formed and people scoring below some specified level are given some form of remediation while those above the cut off move on to a new set of instructional materials. As previously noted, use of scores for decisions such as this involves implicit predictions. For example, it is assumed that those selected to move on will do better on the new material than those assigned to some form of remediation. Other implicit predictions are that students given remedial work will perform better on the test the second time around and will do better on the next segment of instruction after remediation than they would have done without it.

The implicit predictions such as the above can sometimes be supported by empirical evidence. The performance of students before and after remediation can be compared. Experiments could be conducted where some students below the cutoff score on a test are randomly selected to receive remediation while the others move on to new material. The performance of the two groups at a subsequent time could then be compared. The results of such studies would provide evidence of criterion-related validity, albeit not necessarily summarized in terms of a traditional "validity coefficient."

Hambleton and Novick (1973) suggested that the proportion of times

persons above the "qualifying score" on a criterion-referenced test also are above the qualifying score on a "new test" provides an indication of validity. The new test "...might well be derived from performance on the next unit of instruction, or it could be a job-related performance criterion" (Hambleton & Novick, 1972, p. 168). This position is in close harmony with traditional psychometric notions of criterion-related validity. The only differences are in the metric of the scores and the type of summary statistic used. As argued by Hambleton and Novick, a dichotomous score metric is more consistent with some uses of criterion-referenced tests than is the more finely gradated number right score and the proportion of agreement is a more useful statistic for the resulting binary variables than is a product-moment correlation.

Construct Validity

Construct validity has been a controversial topic even in the realm of psychological measurement where it was originated. Its role in educational measurement has been even more dubious. "Construct validity is not usually sought for educational tests, because they are typically thought to be valid on other grounds, namely, on the grounds of content validity" (Messick, 1975, p. 959). This is particularily true of criterion-referenced measures.

Operational definitions are thought of as the key for criterion-referenced tests and no need is seen to invoke constructs in the interpretation of performance (e.g. Osburn, 1968; Harris, Pearlman, & Wilson, in press). But, interpretations of scores on criterion-referenced tests that go beyond operational definitions are commonly made. Competency-based educational programs, for example, rely heavily on criterion-referenced measurement. Yet the very word "competency" implies a construct. Even in very simple contexts the claim of competence or incompetence involves an inference. The claim may be based on a set of well-defined procedures to measure the performance of an individual,

but poor performance does not necessarily imply incompetence.

> "The inference of inability or incompetence from the absence of correct performance requires the elimination of a number of plausible rival hypotheses dealing with motivation, attention, deafness, and so forth. Thus, a report of failure to perform would be valid, but one of inability to perform would not necessarily be valid. The very use of the term inability invokes constructs of attributes and process, whereas a content-valid interpretation would stick to the outcomes" (Messick, 1975, p. 960).

Inferences about competencies need to be supported by the process of construct validation. The need for construct validation is avoided only by avoiding the inferences. In some circumstances such inferences may be avoided altogether, but we think more often than not some inferences will be desired and construct validation will be called for.

It is not possible to provide a simple prescription for establishing construct validity. It is better thought of as a continuing process of marshalling evidence to support or refute inferences and interpretations of test results (Cronbach, 1971; Messick, 1975). The process involves logical analysis as well as a wide variety of possible empirical studies, including for example studies of the effects of experimental interventions.

The process of construct validation has been described most thoroughly in Cronbach's 1971 chapter in Educational Measurement. He illustrated many procedures but noted that "...the procedures cannot be cateloged exhaustively and no guide can tell just how to meet the requirement of hard-headed reasoning from data to conclusions" (p. 484). The emphasis, however, is on a purposeful approach starting "...with a reasonably definite statement of the proposed interpretation" (p. 483). Data which can support or refute the interpretation are collected and the results are used in the refinement of the interpretation.

Instructional theory is far from the stage of providing the type of elaborated nomological net that Cronbach has described. Construct validity cannot be "established" by merely plugging a test into the network and checking on the predictions. The process is both more difficult and less clearly defined than that. On the other hand, it is only through the doing that a network can be developed and refined. In this sense the process of construct validation is just as central to the development of theory as the theory is to the validation of the measures. It is a long term iterative process.

SETTING STANDARDS

Performance standards play a very important and pervasive role in many of the conceptualizations and applications of criterion-referenced measurement. They are fundamental to the frequently cited definition of a criterion-referenced measure provided by Glaser and Nitko (1971). They defined a criterion-referenced test as "...one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). Classification of people as masters or non-masters with regard to a specific domain obviously implies a standard; so does the application of criterion-referenced measures in competency-based programs.

The prominence of standards in applications of criterion-referenced measurement has increased along with the emphasis on competency-based education, and with demands for accountability and for minimum standards for high school graduation. Glass (1976) has shown how the meaning of criterion-referenced measurement has shifted from the time Glaser (1963) originally introduced the term. As noted by Glass, the emphasis in Glaser's original discussion was on determining what an individual can do along a carefully defined and very specific performance continuum. Over the years, however, the emphasis has shifted so that "... the term is now taken to mean tests that relate performance to absolute standards rather than to the performance of others" (Shepard, 1977, p. 3).

Although of fundamental importance, the problem of setting standards has received considerably less systematic attention than other problems such as reliability estimation, determination of test length, or procedures for generating items. This relative lack of attention is unfortunate because a set standard is often the foundation on which an application is based.

Many of the statistical and psychometric formulations which are discussed in other sections of this paper, depend on the acceptance of a standard. If the standard lacks adequate justication the techniques will be of little, if any, value. Glass (1976) put it more strongly. He suggested that the statistical and psychmetric treatments of problems of criterion-referenced measurement represent "...misdirected precision and axiomatization" (p. 36). The prior issues involve the setting and justification of the standard.

The problem of standards is not unique to criterion-referenced measurement. There were standards for graduation or for passing from one grade to the next long before the notion of criterion-referenced measurement was introduced. A standard is implicit, albeit somewhat fuzzier, when children in given schools are reported to be performing a year below grade level. Thus, criterion-referenced measurement did not introduce performance standards to education but it did make them more explicit and has reinforced a tendency to form dichtomies such as master-non-master, competent-incompetent, or pass-fail.

The creation of sharp dichotomies is seen by some as a negative attribute of the criterion-referenced measurement movement. Certainly, if set capriciously, standards are apt to detract from the potential advantages associated with other aspects of criterion-referenced measurement. The National Education Association's Guidelines and Cautions for Considering Criterion-Referenced Testing (1975) caution that "'Minimal comptency' or 'mastery' cut-off points should be viewed with some suspicion" (p. 11). In the discussion of this caution it is argued that "...the setting of such standards is extremely arbitrary." (NEA, 1975, p. 11). The concern that standards are often arbitrary and apt to be a negative aspect of criterion-referenced measurement was well-articulated by Dyer (1977) who wrote:

> "Now there is a new term, criterion-referenced testing, which all lovers of new jargon proudly hail as the great new device for mitigating the ego threat commonly attributed to odious comparisons inherent in norm-referenced testing. ...It seems to me rather ironic that when the notion of norms first found its way into the vocabulary of testing, it represented an attempt to get away from the odium of the pass-fail criterion. Norms, properly considered were to be neutral indicators of relative position; they were to say nothing whatever about passing or failing, winning or losing. Yet the notion of passing or failing is precisely the one that, in a painfully refined and exagerated form, the idea of criterion-referenced testing has brought back to the language of measurement. Just last month, one of our local papers was reporting the percentage of "passes" that each of various school districts had achieved on the reading tests administered in New Jersey this year. And in this case the criterion is the familiar old fashioned one of getting right answers on 65 per cent of the items." (p. 18-19).

We desperately need to do better than arbitrarily setting a standard at 65 per cent or 80 percent for no better reason than habit. Investigations of various methods of setting standards and of their effects are needed. This is largely unexplored territory. Some recent work (e.g., Block, 1972; Huynh, 1976b; Meskaukas, 1976; Millman, 1973; Jaeger, 1976; Shepard, 1976) may provide some guidance on how to proceed. But, there are flaws in all of the approaches that have been suggested (see, for example, Burton, 1976; Glass, 1976).

Judgmental Nature of Standard Setting

The above introductory comments are not intended to imply that standards are always set without thought or effort. Nor, are they meant to imply an absence of suggested procedures for establishing standards. Procedures have been suggested and there are examples where great care and effort has gone into the establishment of standards. Unfortunately such cases are the exception rather than the rule and even the best examples have some arbitrariness.

Millman (1973) reviewed a variety of procedures for setting standards and classified them according to whether they were dependent on (1) a comparison to the "performance of others," (2) an analysis of "item content,"

(3) an evaluation of their "educational consequences," (4) an evaluation of their "psychological and educational costs," and/or (5) adjustments for "errors due to guessing and item sampling." More recently, Glass (1976), suggested a system of six categories, five of which roughly correspond to Millman's. The additional category involves a process of "bootstrapping on other criterion scores". These categories are not mutually exclusive and hybrid procedures cutting across categories can be imagined. Regardless of category or category system, all the procedures share one important feature. They all involve judgment. They differ in terms of the sources of information and the type and extent, if any, of empirical evidence that is needed in order to make the judgment. But, "None of the procedures eliminates the need for judgment." (Millman, 1973, p. 206).

The pre-eminence of human judgement in standard setting is widely acknowledged. For instance, Messick (1975) claimed that "...all procedures for establishing performance standards require judgment at some point" (p. 957) and according to Jaeger (1976):

> "All standard-setting is judgmental. No amount of data collection, data analysis, and model building can replace the ultimate judgmental act of deciding which performances or which levels of performance are meritorious or acceptable and which are unacceptable or inadequate." (p. 22).

This does not imply that empirical data are irrelevant to the process. On the contrary, they may be of great value in facilitating judgment making, "...but they cannot be used to ferret out standards as if they existed independently of human opinions and values" (Shepard, 1976, p. 29). Thus, the procedures that are briefly described below are best considered as aids to judgment and not substitutes for it. Methods for making the judgmental process more systematic and for evaluating the process are also needed.

Types of Models

Meskaukas (1976) distinguished between two broad classes of mastery models which he refers to as "continuum models" and "state models." In a continuum model, performance levels are conceived of as varying along a continuum, but the continuum is dichotomized into two regions Persons in the upper region are referred to as masters and those in the lower region as non-masters. The dichotomization is considered useful, or essential, for certain educational decisions. There is no "natural" or obviously preferred way of forming a dichotomy when it is recognized that the performance of interest varies along a continuum, however.

In a "state model," learning is conceptualized in an all-or-none fashion. Students are either masters and therefore can perform the task, or they are not masters and cannot perform it. They either know the answers to all the items in the content domain or they don't know any of them.

In state models, the setting of standards is natural. The measurement task is simply to provide the means of deciding which of two states a person is in at a particular point in time. In a continuum model, on the other hand, the task is two-fold. First, one must decide how to dichotomize the continuum, then the measure must be used to decide in which region of the continuum the person belongs.

Meskaukas (1976) reviews procedures for setting cutting scores on tests and obtaining mastery decision rules for the two types of models. Our main concern will be with procedures for continuum models. There are two reasons for this focus: (1) only continuum models require both the judgmental task of setting a standard as well as the procedural task of specifying the decision rules, and (2) it is our contention that relatively

few areas of achievement that are broad enough to correspond to important content domains are truely all-or-none tasks. Meskaukas argued that "... a great deal of what is learned, particularly in situations where errorless replication is required, follows...[the state] model" (1976, p. 155). But errorless replication is only one half of the coin. A state model would also require that the only other possibility would be consistently replicated errors. At least while learning a content domain, some intermediate outcomes are usually possible. In some situations, arguments for a state model may actually be arguments that the mastery region should contain only perfect performance.

If the domain consists of a single addition problem, say 2 + 2, then a state model is most appealing. If the domain is expanded to include the addition of all possible pairs of one-digit numbers, however, knowing all the answers or knowing none of them are no longer the only logical possibilities. We concur with Millman's (1974) conclusion that "Because these models assume that partial knowledge or skill does not exist, they are considered unrealistic...." (P. 355).

It has been suggested that domains should be subdivided to the point that the items they contain are so homogeneous that a state model is reasonable (Macready & Merwin, 1973; Macready & Dayton, 1977). If this is done, then setting the criterion level is no longer an issue, and the problem then is reduced to assessing the probability that a person is a master and models for doing this have been developed by Macready and Dayton. The extent to which a state model can provide a useful approximation for domains that are not too narrow to be of interest is somewhat problematic, however.

Item Content

Domain specifications and/or item content may be the primary source of information used in judging the level at which to set standards. The formal procedures reviewed by Millman (1973) for using item content to set standards involve the analysis of the individual items on a specific test. These approaches ignore the issue of setting a standard for the domain and move directing to setting a passing score for the specific sample of items that comprise the test.

The distinction that is made in discussions of test length (see below) between "criterion level" and "passing score" are often obscured in the actual setting of standards. Criterion level refers to the minimum proportion, π_0, of all items in the domain that are required for a person to be a "master", "acceptable", or "competent". The passing score, n_0, on the other hand, is the score on a particular test that is required to be classified as a master. Since the items on the test do no exhaust the domain, classification errors will be made regardless of the type of correspondence between the passing score and the criterion level. Frequently the distinction between the criterion level and the passing score is ignored. Procedures are used to establish n_0 and the π_0 is implicitly assumed to be equal to n_0 divided by the total number of items on the test, n. The two are conceptually distinct, however, and it is not necessarily desirable to set n_0 equal to n times π_0.

Ideally, standard setting procedures would be used to establish π_0 and then n_0 would be determined in view, not only of π_0, but also of the disutility associated with false-positive and false-negative errors. None of the item content procedures reviewed by Millman do this. Instead, n_0 is set directly by reviewing the items on a test and in one form or another, deciding the number of items that a "minimally acceptable person" should get right. The approaches vary in detail (see Millman, 1973 for

more description) and may involve a direct estimate of n_o (e.g., Science Research Associates, 1966) or indirect estimates of n_o by summing estimates of "... the proportion of minimally acceptable persons who would answer each item correctly" (Angoff, 1971, p. 515). Some would combine judgments of item importance and expected proportion of minimally qualified persons passing each item (e.g., Ebel, 1974; Educational Testing Service, 1976) Yet another approach (Nedelsky, 1954) involves the elimination of distractors that the minimally qualified person should be able to eliminate.

The role of judgment is obvious in all these methods. More needs to be known about the sensitivity of the results of these methods to the nature and variety of audiences involved in the judgments, to the nature and training for the task, and to the availability and nature of empirical data on the performance of various groups. Evidence regarding the inter- and intrajudge consistency is also to be desired.

The different procedures can be expected to yield different standards. (Andrews & Hecht, 1976; Glass, 1976). Such differences reveal that there is an arbitrariness to the standards that may result from a given procedure. Furthermore, the very notion of "minimal competence, on which the procedures depend, is questionable. After a consideration of the logical and psychological bases for the concept of minimal competence, Glass concluded that "the idea of minimal competence is bad logic and even worse psychology" (1976, p.32). For any non-trivally low "minimum" required for some other activity exception can surely be found. Such exceptions violate the logical basis for the label "minimum".

Educational Consequences

The educational consequences of setting standards at various levels may be used in place of, or in addition to, item content to set standards (Millman, 1973). For example, in a mastery based instructional program, the effect on learning in subsequent units of the program could be used as the primary basis for determining the desired passing score. If

experience shows that students with scores on a unit test of less than 85 percent have undue difficulty while those with higher scores generally progress well in the next unit, then 85 percent would be selected as the passing score.

Empirically investigating the educational consequences of various possible passing scores is a non-trivial task; one which has rarely been attempted. Block's (1972) investigation is relatively unique in this regard. He compared the subsequent performance of randomly formed groups of students that were required to meet various performance levels on a test of matrix algebra before moving to the next segment of instruction. The "best" cutting score was found to vary depending on the emphasis that was placed on cognitive or affective criteria in the instructional segments following the matrix algebra test. Judgement is required to determine the way in which to weigh the criteria. Indeed, judgement is required to decide what criteria should be considered. There is bound to be some arbitrariness in these judgements.

Huynh (1976b) has developed a mathematical model that provides a way of setting the standard in terms of future performance on what he calls a referral task. The probability of success on the referral task, the distribution of true proportion correct, the loss associated with a false positive decision, and the loss associated with a false negative decision are all used in the determination of the optimal criterion level, if one exists. The optimal criterion level is in turn used to solve for the cutting score on the test.

Huynh's model provides a useful conceptualization. But, it is apt to prove very demanding in practice. Setting the loss associated with each type of error requires the completion of a difficult judgmental task. Obtaining the evidence needed to establish the probability of success on

the referal task for each unit of an individualized instructional program is formidable, if not an overwhelming task. It is still more difficult to imagine how one could obtain the empirical evidence that would support the claim that a given minimal standard for high school graduation was optimal in terms of future consequences of various possible standards.

Millman (1973) suggested that "In the absence of data about the educational consequences ..." that "... a logical analysis of the subject matter and the extent of the instructional system..." (p. 209) be used as the basis for setting the passing score. Relatively high passing scores would be required where future learning was seen as clearly dependent upon the skills and knowledge of a unit, while lower passing scores would be acceptable where there was not such a clear dependency. This approach obviously requires complex judgments the difficulty of which undoubtedly varies from area to area. Where more major decisions, such as those required to set standards for high school graduation, are to be made, the determination of necessary prerequisites becomes problematic since we lack clear answers to a prior question; i.e., prerequisites for what?

Comparative Data

The idea that comparative data on the performance of others might be used in the setting of standards would seem to be an anathema to a proponent of criterion-referenced measurement. Although possibly useful for selection (Millman, 1973), the setting of a passing score so that a fixed percent of the examinees pass would generally be eschewed by a criterion-referenced measurement devotee. Indeed, getting away from the comparison of performance of an individual to that of others is precisely what is often seen as one, if not the main, advantage of criterion-referenced measurement. But, it is apt to be impractical to set standards and apply them without regard for their normative implications.

The following example of an actual experience of a school district illustrates the type of problem that can result if normative results are completely ignored. The staff of a school district was given the mandate to construct a test that would be used to establish minimum standards for high school graduation. The teachers and other professionals within the school district devoted considerable time and energy to defining the domains to be covered, constructing the test items and making the judgments necessary to set the standards. The task was taken seriously and a conscientious effort was made to set a passing score in reading and one in math that represented minimal performance levels for high school graduates. A tryout of the tests revealed, however, that application of the passing scores would result in the failure of approximately 25 percent of the students on the reading test and 45 percent on the math test. Would failure rates such as these be educationally or socially desirable? Would they be politically feasible? We think not.

It might be argued that the teachers making the judgments must have used inappropriately stringent standards. But, why would others be expected to make more appropriate judgments? It might also be argued that, if the tests were given early enough, they could identify potential failures who could be given remedial instruction before they are given alternate forms of the tests. It seems unreasonable, however, to expect that a brief period of remediation is likely to teach students the minimal essentials when that couldn't be accomplished in the previous eleven years of schooling. Certainly, our experience with the results of compensatory education would hardly make us sanguine with regard to this possibility.

Another illustration of the importance of nomative feedback comes from the Michigan Assessment Program. The assessment provided information on the accomplishment of "minimal objectives" by school districts in the state. In their critique of Michigan Assessment

Program, House, Rivers, and Stufflebeam (1974) argued that there is "... considerable reason to believe that the objectives are not minimal" (p. 6) and cite in support of this the result that not a single district in the state was achieving "minimal objectives" as they were defined. Such an outcome defies the common sense meaning of the word "minimal", to say nothing of a strict interpretation of the word.

Knowing that 25 percent or 45 percent of the students would fail to meet a standard identifies a problem but doesn't tell us what to do about it either in terms of educational changes or in changes in the standard. It is information, however, that can hardly be ignored in the overall judgmental process of setting standards.

Threats to Validity of Inferences Based on Standards

Jaeger (1976) discussed several approaches to setting standards ranging from direct methods such as the item content procedure described above to "distal" methods involving dependence on normative data or the relationship of a test to an external measure. For each method he reviewed the ways in which the validity of inferences based on the standard might be threatened. A person who meets the standard is considered a master and should be able to do certain things. The inference that the person is competent to perform certain tasks may be erroneous for a variety of reasons.

A total of thirteen threats to validity were identified by Jaeger. Different methods of setting standards are subject to different threats. But, seven threats to validity are present regardless of the method used to set standards. Four of these threats are present when the test results are used to make inferences about domain performance. An inference about domain performance is made, for example, when the score on a test is used to infer that a person is above the criterion level for the domain from which the test items are sampled. The other three threats to validity,

that are present regardless of the method used to set standards, concern inferences about ultimate criteria. This latter type of inference will be considered more fully below.

According to Jaeger (1976), when inferences are limited to statements about the performance domain from which the test items are sampled, one is always faced with at least the following four threats to validity:

> "Bias in setting domain standard [i.e., criterion level] due to inadequate domain definition."
>
> "Random error among judges who set domain standard."
>
> "Inappropriateness of item sampling procedures: bias error in sample standard." (i.e., passing score).
>
> "Inadequate item sample size: random error in sample standard" (p. 26).

With rare exceptions, such as the domain of all addition problems with two addends of one digit each, the first and third threats to validity are apt to be particularly serious. As was discussed in previous sections of this paper, adequate domain definition is crucial. It is required not only for adequate test construction, i.e., construction such that the test is representative of the domain, but also to make appropriate judgments about the desired criterion level.

Although the threats to validity of inference about domain performance are serious, they pale by comparison to those for inferences about ultimate criteria and the latter inferences are usually the ones of primary interest. In isolation, inferences about domain performance "... are often uninteresting and insufficient" (Jaeger, 1976 p. 24). When someone meets the minimum standards on a test for graduation from high school, the types of inference might be limited to ones about performance level of the individual in the well-defined domains represented by the tests. The desire for this limited type of inference, however, was not the force that has led to the demands

for minimum standards in states throughout the country. Inferences about an individual's competency for various "life roles" are also desired. For example, Kohr (1977) described a bill before the Pennsylvania state legislature that deals with minimum competency requirements. These would include, among other things, evidence that a student had minimum competency to "... 'understand and perform personal finance and consumer tasks including understanding consumer finance; computing interest rates, purchasing insurance, completing personal tax forms, knowing the basis of property and other taxes and comprehending residential leases and purchasing agreements'" (from Pennsylvania House Bill 770, as quoted by Kohr, 1977, p. 2). The inference that someone understands consumer finance or comprehends residential leases goes well beyond the inference that he or she can perform X percent of the items in a domain. Not only that, but the motivating force actually comes from a concern about future behavior; i.e., how the person will perform as an adult. Concern about future performance is not unusual. In fact, "Often our interest in current performance only substitutes for our true interest in later performance, perhaps years later" (Jaeger, 1976, p. 24). In short, inferences about "ultimate criteria" are often desired and these inferences go well beyond domain performance in terms of scope, time and the degree to which they can be translated into observable behavio .

For inferences about ultimate criteria that are based upon set standards, Jaeger identified the following three threats to validity:

> "Bias error due to invalidity of domain definition."
>
> "Error due to inconsistency of domain-criterion relationship."
>
> "Bias error due to invalidity of model for domain-criterion relationship." (1976, p. 26).

Combined, the threats to validity for the two types of inference produce a list that "... is discouraging, if not mind-boggling, and our knowledge of the magnitude of errors and the severity of such validity threats is extremely limited." (Jaeger, 1976, p. 25).

The process of setting standards and of making inferences based on those standards lacks a firm foundation. This creates problems that are apt to be of more than academic interest. The setting of standards and their use for certification has already been at issue in some recent court cases (e.g., U.S. v. State of South Carolina, 1977). The apparent trend toward increased use of tests to establish that standards of minimum competency have been met for various types of certification is apt to lead to more cases involving the use of standards. Typical current practice is not likely to provide a basis that most standard setters would be comfortable defending in court.

Empirical investigations involving real domains, inferences, and judges could provide answers to some important questions about the threats to validity (see, Jaeger, 1976). The results of these investigations coupled with considerable thought might lead to a better theoretical foundation as was suggested by Jaeger. But, a concerted effort would be required and there are few signs that such an effort is underway, though calls for the needed research are not new (e.g., Airasian & Madaus, 1972; Quirk, 1974).

Are Standards Necessary?

Glass (1976) has argued that all existing methods of setting standards are so seriously flawed that they are apt to do more harm than good.. As has been noted several times, there is an arbitrariness in all of the methods of setting standards. Glass would want to avoid this arbitrariness by not using standards. He suggested that rather than trying to set absolute standards the focus should be on change. Improved performance is good and a decline in performance is bad. Absolute standards are viewed as unnecessary and, because of their arbitrariness, best avoided.

We agree with Glass' conclusion that every method of setting standards involves arbitrariness at some stage. We also endorse his suggestion that

comparisons over time have advantages over absolute judgements. We doubt, however, that all demands for standards can be so easily sidestepped. Many types of educational actions require something besides evidence of improvement. Knowledge that the performance of a student is better now than at sometime in the past is useful but it doesn't answer questions about whether the student is prepared to move on or still should receive some form of remediation.

Comparing the performance of graduates to that of previous graduates is useful. Agreement that a decline is bad, could probably be obtained from most people.* Unless all forms of certification were eliminated, however, there is still a need for a standard to determine whether the performance is sufficient to receive the certification. The standard will admittedly be based on some arbitrariness. It will not be an absolute. This does not imply, however, that a standard is worthless.

People will be categorized and differential actions taken on the basis of those categories with or without standards that are systematically set. The categorization will involve human judgement and a degree of arbitrariness with or without systematic methods. But, the more systematic methods have certain advantages. They are more explicit and can therefore more readily be made public and be subjected to debate. They need not be viewed as fixed once they are established. Indeed, as will be suggested below, part of the systematic procedure should provide for frequent review and revision.

Standards, set by any method, may be arbitrary, but they need not be capricious. Systematic procedures can reduce the chances of capriousness. Glass (1976) concluded that "setting performance standards on tests and exercises by known methods is a waste of time" (p. 49). We disagree. The setting of standards is an important problem, one that deserves more

* The recent controversy over the meaning and importance of the decline in scores on the Scholastic Aptitude Test suggests that not everyone would even agree that a decline is bad.

time and effort that it normally receives. The methods all have defects, but are better than nothing because in this case, "nothing" really means hidden or unknown standards.

Interim Suggestions

In the absence of an adequate theory on which standard setting procedures can be developed, only tentative guidelines and opinions can be offered. What we judge to be a very reasonable set of recommendations has recently been offered by Shepard (1976). She suggested that standard setting should be considered an iterative process and one that involves various audiences. The involvement of various audiences in the task allows for differences in points of view of these audiences. It gives explicit recognition to the complexity of setting a single standard when different standards would be selected by different groups of judges. While this may complicate the process when the groups are found to differ substantially it is far preferable to the misleading simplicity of using only one relevant audience to set a standard.

Shepard's suggestion that the process be an iterative one permits adjustments in the judgments based on the accumulation of information over time. Thus, in the example used above where the application of graduation standards would have resulted in an unacceptably high failure rate, this information would be fed back to the judges who might want to modify their judgments in light of the additional information.

Another of Shepard's recommendations is that normative data ought to be provided to judges for use in their deliberations. The norms are not to be used to set the standards; merely to inform the judges.

Shepard's remaining recommendation is more relevant for assessment or accountability systems than for mastery based instruction or the certification of individuals. This recommendation is that improvement rather than current performance be the standard that is employed.

Shepard's suggestions are sensible. They do not avoid the dependence on human judgment, nor could they. Neither do they eliminate Jaeger's threats to validity of the inferences based on the use of standards. We think, however, that they provide the broad guidelines within which defensible methodologies for setting standards may be developed. We also concur with Conway's (1976) conclusion that "Unless standards are established by some defensible methodology which involves careful human judgment, they will not serve their intended purposes nor will they stand up against the careful scrutiny of those who doubt their validity"(p. 35).

TEST LENGTH AND PASSING SCORES FOR MASTERY DECISIONS

In classical test theory, considerations of test length have generally been focused on the effects of length on reliability and validity. The well-known Spearman-Brown formula is a prime example of this type of analysis (see, for example, Lord & Novick, 1968, pp. 103-119). More complex analyses involving the length of tests that optimize properties of a battery of tests have been considered (e.g., Woodbury & Lord, 1956; Woodbury & Novick, 1968 pp. 119-125) but the emphasis on reliability and validity coefficients is the same. Discussions of test length for criterion-referenced measures have addressed rather different issues, however.

Criterion-referenced measures are commonly used to make dichotomous decisions. Persons scoring above some specified level are typically labelled masters and those scoring below that level are labeled non-masters. Such classifications, sometimes involving different labels, occur in several contexts, but the prototypical example is individualized instruction where masters move on to more advanced instructional materials while non-masters receive additional instruction on the domain covered by the test. In such a context, the main question of test length is concerned with the number of test items needed to make the mastery–non-mastery decisions with satisfactory accuracy.

The concept of an item domain with the associated notion of a test as a sample from that domain have been used in addressing the problem of test length for mastery decisions. In the context of a well-defined domain and a test as a sample of items from that domain, ..."the test length problem is determining the size of such a sample... needed to acquire an estimate having a specified level of accuracy" (Millman, 1973, p. 211).

Binomial Model: Scores Conditional on True Proportion

Millman (1972, 1973) proposed a binomial model for infinite or very large domains of items. The proportion of all items in a domain that an individual could answer correctly, π, would ideally be compared to a prespecified criterion level, π_0, and the individual would be considered a master if $\pi \geq \pi_0$ while a person with $\pi < \pi_0$ would be considered a non-master. In practice, of course, not all of the items in the domain can be administered. Indeed it would be desirable to administer a relatively small number of items and use an estimate of π to make a decision for an individual.

With the binomial model used by Millman, it is quite easy to calculate the probability that a person with a given π would equal or exceed any specified passing score on a test of length n (where the n items are assumed to be randomly sampled from the domain). Millman (1972) used this approach to generate tables which reported the probability of being misclassified as a function of: 1) the criterion level, π_0, 2) the proportion of items in the domain that would be correctly answered by a person if he/she were given the entire domain of items (i.e., the true proportion correct), π, 3) the number of items, n, and 4) the passing score, n_0. For example, with $\pi_0 = .80$, $\pi = .70$, $n = 10$, and $n_0 = 8$, the probability of being misclassified is .38. In other words, students with $\pi = .70$ would be expected to get 8, 9, or 10 of 10 items right, and therefore be misclassified as a master about 38% of the time. The probability of a misclassification increases as π approaches π_0. Such an increase in the probability of a misclassification is to be expected since it is obviously more difficult to determine if a person with $\pi = .79$ is above or below a criterion level of .80 than it is for a person with $\pi = .20$. The misclassifications for people with π's close to π_0 are also apt to be considered less serious than are misclassifications for persons with π's far removed from π_0. This distinction between the seriousness of misclassifications

according to the size of the discrepancy between π and π_0 is important to procedures considered below involving an indifference region.

In general, the probabilities of misclassification reported by Millman decrease as the length of the test and the passing score increase proportionately. The pattern is somewhat irregular, however, since only integer values of n and n_0 can be meaningfully considered. A striking aspect of Millman's table is that it suggests that much longer tests are needed than are typically used to make mastery decisions in individualized instruction programs unless one is willing to accept relatively high probabilities of misclassification for people with π's within, say, .10 of π_0. For example, with $\pi_0 = .80$ even a relatively long test of 25 items with a passing score of 20 would be expected to misclassify someone with $\pi = .70$ as a master about one-fifth of the time (see Millman, 1973, Table 1).

Use of an Indifference Region

One problem in applying Millman's tables to arrive at a decision about test length and passing score is that the probabilities are conditional upon an unobservable quantity, π. In practice, one would need to know something about the distribution of the π's in order to have an idea of the number and types of misclassifications to be expected. Fhaner (1974) presented a procedure which is closely related to the Millman approach but does not require prior knowledge of π.

Like Millman, Fhaner results are based on a binomial model. Rather than calculating a probability of misclassification conditional on π, however, Fhaner suggested calculating probabilities of misclassification conditional on the end points of an indifference region. Thus, an interval from π_1 to π_2 would be identified such that $\pi_1 \leq \pi_0 \leq \pi_2$. The interval would be selected so that it would be considered unimportant whether a person with a value of π within the interval was considered a master or a non-master. Given π_1 and π_2

the maximum probabilities of the two types of misclassification, i.e., classified as master when $\pi < \pi_1$ or classified as a non-master when $\pi > \pi_2$, would need to be specified. The maximum probability of misclassifying someone as a master is called α and the maximum probability of misclassifying someone as a non-master is called β.

Once π_1, π_2, α and β have been specified, Fhaner's procedure would then identify the test length and associated passing score, n_0, such that the probability that the observed number right is greater than n_0, given $\pi = \pi_1$, is less than or equal α and the probability that the observed number right is less than or equal n_0 given $\pi = \pi_2$, is less than or equal β.

For large n, Fhaner suggested using a normal approximation to the binomial distribution. But, his results show that the normal approximation tends to underestimate the required test length. As noted by Wilcox (1976), the approximation is not really necessary since accurate values of n and n_0 can be obtained using the incomplete beta function in place of the binomial distribution.

Fhaner's results clearly indicate the need for quite long tests for small indifference regions and small values of α and β. For example, with $\pi_0 = .75$, an indifference region from .70 to .80, and $\alpha = \beta = .10$, a test of over 100 items would be required. Required test length falls rapidly as the length of the indifference region is increased. For example, if the indifference region was changed to .65 — .85, while keeping $\pi_0 = .75$ and $\alpha = \beta = .10$ in the above illustration, a test of 31 items would be required. An indifference region of .60 to .90 would require only 15 items.

Wilcox (1976) extended the work of Fhaner by providing a procedure that will always yield a recommended test length and associated passing score. The recommended values are selected such that the probability of a correct decision is at least P*, which is some preassigned value between .5 and 1.0. Tabled values of n and n_0 along with their associated minimum probability of

a correct decision are reported for indifference regions of length .10 and .20 and for criterion levels of .7, .75, .8 and .85. It is clear from Wilcox's results that with the shorter indifference region, relatively long tests are required if a high value of P* is to be achieved. For example, more than 20 items would be required to achieve a P* of even .75 when the indifference region is only of .10 in length.

Bayesian Procedures

As noted by Novick and Lewis (1974), the misclassification problem with short tests is quite serious. Unless one is willing to accept relatively large probabilities of misclassification and/or accept a relatively large indifference region, test lengths are required that are greater than what is apt to be considered practically feasible. One approach to solving this apparent dilemma is to use Bayesian procedures to take advantage of prior knowledge about the distribution of π. The Bayesian approach also has the advantage that it provides estimates of "...the probability that a student's true level of functioning [π] exceeds the specified criterion score" rather than "...the probability that a student will attain a particular test score, given his true level..." (Novick & Lewis, 1974, p. 142).

The Bayesian procedure proposed by Novick and Lewis is based on the binomial model. As has been found useful in other applications involving the binomial model (see, for example, Novick & Jackson, 1974), Novick and Lewis proposed that prior beliefs about π be characterized by a Beta distribution. In the absence of any prior information about the distribution of π, the parameters of the prior Beta distribution would both be set equal to 1.0, which is equivalent to saying that all equal intervals from .0 to 1.0 are equally likely for π. With this uniform prior distribution, the Bayesian results are quite similar to those that would be obtained using the Millman (1972, 1973) or Fhaner (1974) procedures. Novick and Lewis argue, however, that "...it

will seldom be the case that our [prior] knowledge of a student's true level is adequately described by a uniform distribution". (1974, p. 144).

With stronger prior knowledge, the Bayesian procedures presented by Novick and Lewis can yield substantial savings in test length without compromising misclassification probabilities. For example, with a uniform prior distribution the probability that π is greater than $\pi_0 = .80$ given a score of 7 or 8 correct on an 8-item test is .56. In other words, the probability that people with scores equal or above the passing score are actually masters is only a little over a half. On the other hand, if prior knowledge is assumed "...to be roughly equivalent to that based on a score from a 12 item test" and the "...prior probability that a student is functioning above a criterion level of .80...[is] approximately .75" (Novick & Lewis, 1974, p. 144) then the corresponding figure is .81. That is, given the above prior knowledge, the probability that a person with a score of 7 or 8 correct on an 8-item test is in fact a master is .81 rather than the .56 estimated in the absence of prior knowledge.

Novick and Lewis considered 20 different prior distributions, the combinations of 4 different effective prior test lengths (8, 10, 12 and 15), and 5 different mean values for the prior distributions of π (.7, .75, .8, .85, and .9). They also considered 4 different criterion levels (.7, .75, .8 and .85) and 4 ratios of loss associated with a false-positive error to the loss associated with a false-negative error. The latter figure provides explicit recognition of the possibility that the error of misclassifying someone as a master may be more serious than the error of misclassifying someone as a non-master.

Using the above specifications, Novick and Lewis identify test lengths and associated passing scores that are recommended for each situation. Test lengths greater than 25 items were judged impractical and were not

considered. Recommended test lengths and passing scores were found to be sufficiently uniform over the four prior distributions with the same mean that a single general recommendation of test length and passing score was made. The criterion level, the ratio of losses associated with the two types of errors, and the mean of the prior distribution were found to have important effects on test length and pass score recommendations, however.

As noted by Novick and Lewis, there is a trade off between instructional time and testing time. Shorter tests are needed in situations where the mean of the prior distribution is higher, which would presumably be more likely with longer instructional time. They recommend that instruction be carefully monitored such that "...the expected group performance will be just slightly higher than the specified criterion level." (p. 157). They also recommend that a criterion level less than .85, preferably around .75, be used. Finally, they suggest that the instruction be planned such that the seriousness of a false-positive error is not much greater than that of a false-negative error.

A variety of other Bayesian models might be considered. For example, as suggested by Hambleton, Swaminathan, Algina and Coulson (1975), a compound-binomial model might be used in place of a binomial one. It would also be possible to consider loss functions other than the threshold loss function used by Novick and Lewis. Indeed, three such possibilities, two of which allow for graduation in the loss as a function of the size of the difference between π and π_0, were suggested by Novick and Lewis. An indifference region such as the one used by Fhaner could also be incorporated in a Bayesian model. Regardless of the specific Bayesian model, however, the choice of the prior distribution is a key and controversial aspect of the approach.

The Bayesian model has distinct advantages provided the model holds. It can yield a given accuracy of classification with a shorter test than the

non-Bayesian models. Or, for a given test length, the Bayesian model can provide smaller errors. But, "If the Bayesian prior distribution assumptions don't hold, even larger errors can occur." (Jaeger, 1976, p. 24).

The stipulation of a prior distribution must be done with care. On first consideration, the problem of deciding how strong the prior knowledge and what the expected value of the prior distribution should appear to be formidable tasks. The decision about the strength of the prior information amounts to deciding that prior knowledge is roughly equivalent to the results obtained from a test of some specified length. Novick and Lewis provide several references to aids that should help in the specification of the prior distribution. In addition, it is fortunate that the results obtained by Novick and Lewis are relatively insensitive to the strength of the prior knowledge over a range equivalent to 8 to 15 items. The expected value of the prior distribution was found to be important. Historical data, however, may enable the decision maker to specify the expected value of the prior distribution with somewhat more confidence than he or she can specify the strength of the prior knowledge.

Alternative Error Models

All of the approaches to test length considered above are based on the binomial model. Although the binomial model has been found to provide a useful approximation to observed test score distributions, the compound binomial model may provide a better approximation (see Lord & Novick, 1968, Chapter 23). Recently, Wilcox (in press, a) has presented a general approach to the test length problem that includes the binomial and compound binomial models as special cases. He also generalized to situations where the probability of correctly classifying k examinees is considered simultaneously. When specialized to the binomial model and to k = 1 (i.e., when the probability of a correct decision is considered for a single examinee) Wilcox's approach is

equivalent to that presented by Fhaner (1974) and elaborated by Wilcox (1976). When k is greater than one - the results provide test lengths required to have a specified minimum probability of correctly classifying all k examinees simultaneously.

An encouraging feature of Wilcox's results is that the test length requirements based on the binomial model are apparently conservative. Thus, if a binomial model is assumed where the more complicated compound binomial model would be more appropriate the test length required for a specified minimum probability of a correct decision would generally be expected to satisfy an even more stringent minimum probability under the more appropriate model.

Another notable feature of Wilcox's results is that when one is concerned with the probability of correctly classifying all $k > 1$ examinees, rather than considering a single examinee at a time, very long tests would be required to achieve even modest minimum probabilities. This result may be of little concern in the context of individualized instruction where it may be most appropriate to consider each student as a single examinee for which some specified minimum probability of a correct decision is desired. Thus, the results for $k > 1$ may be of greater theoretical than practical importance. Nonetheless, it is sobering to recognize that even for a test of, say, $n = 25$ items, the minimum probability of correctly classifying all the members of a group of k examinees will generally become very small as k increases. For example, Wilcox (in press, a) reported a typical situation where the minimum probability of a correct decision for all examinees would be about .83 if $k = 1$, .39 if $k = 5$, and only .02 if $k = 20$.

Likelihood of False-Positives and False Negatives

Although not focused on test length, per se, another recent paper by Wilcox (in press, b) has some implications for test length. In the latter paper Wilcox presented procedures for estimating the likelihood of making a false-positive decision, α, and that of making a false-negative decision, β. The

likelihood of making either type of misclassification is a function of the prior distribution of π and the conditional distriubtion of the number of correct answers given π and the test length. A beta distribution was assumed for π and either a binomial or compound binomial distribution was assumed for the conditional distribution of the observed number right score given π. Several estimation procedures were developed by Wilcox and evaluated using Monte Carlo techniques.

Estimates of α and β could be obtained using the procedures developed by Wilcox for various values of n and n_0. The desired test length and associated passing score could then be selected from among those which yielded acceptable estimated values of both α and β. A small illustration of this approach was presented by Wilcox.

All-or-None Mastery Models

If an all-or-none model of mastery is assumed, then the determination of test length can be accomplished using procedures presented by Macready and Dayton (1977). In a strict all-or-none model, only a one-item test would be required since masters would always respond correctly and non-masters incorrectly. The Macready and Dayton models allow for masters missing items due to short-term "forgetting" and non-masters to respond correctly by guessing. Given these possibilities, more than one item would typically be required.

Tables of required test length based on their all-or-none mastery model have been prepared by Macready and Dayton. To use these Tables one needs to specify (1) "the proportion of non-masters in the population," (2) the relative costs of false positive and false negative errors, (3) the maximum acceptable error rate, (4) the probability that a master will respond incorrectly to an item, and (5) the probability that a non-master will guess the correct answer (Macready & Dayton, 1977, pp. 108 & 112).

Sequential Decisions

As previously noted, it is generally more difficult to make a decision for someone with a true proportion correct close to the criterion level than it is for someone far removed from it. Savings in average test length may be achieved for a given overall accuracy level by using sequential procedures. After each item a decision would be made to classify the person as a master or as a non-master or to continue testing. The potential utility of such a sequential procedure is discussed at length by Cronbach and Gleser (1965).

Dichotomous classification using sequential procedures developed by Wald (1947) were compared to classifications using a constant number of items for everyone by Linn, Rock and Cleary (1972). They found that sequential procedures could result in an average savings of about 50% in the test length for a given level of accuracy. Green (1970) reported a similar savings based on his theoretical analysis using a domain of equivalent items. A successful application of sequential procedures in an individualized instruction context was reported by Ferguson (1971).

Novick and Lewis (1974) noted that Bayesian procedures may be used for sequential decisions without any added complications. "Items can simply be administered until it is clear that a student will definitely, or cannot possibly attain the minimum advancement [passing] score. Thus, with a minimum advancement score of 8/10, testing can cease as soon as eight successes or three failures are observed" (Novick & Lewis, 1974, p. 157).

Summary

The issue of test length for criterion-referenced measures has been defined as a problem of deciding on the test length and passing score required to make sufficiently accurate mastery decisions. Both Bayesian (Novick & Lewis, 1974) and non-Bayesian (e.g., Millman, 1972, 1973; Fhanér, 1974) have

been proposed. The non-Bayesian results suggest that longer tests than have typically been used in practice may be required unless large probabilities of misclassification and/or large indifference regions are acceptable. The Bayesian approach suggested by Novick and Lewis has the potential advantage that prior knowledge about the distribution of true proportion correct scores of examinees may make it possible to use shorter tests while still achieving acceptable levels of accuracy of decisions. Probably the main disadvantage of the Bayesian approach is the need to carefully specify the prior distribution which may be difficult and may prove controversial.

The recent work of Wilcox (in press, a, b) has extended the test length analyses in several ways. He provided for error models other than the simple binomial model thereby allowing for possibly more accurate approximations to observed test score distributions. He also provided a generalized model for considering $k \geq 1$ examinees simultaneously. Finally, he has provided an alternative approach to determining test length based on the likelihood of the two possible misclassifications. The latter approach should prove useful when one is concerned about characterizing a group of examinees rather than focusing on a single examinee and determining the conditional probability of correctly classifying that person given his or her observed score.

Where it is feasible through self-scoring tests, or possibly computer administered tests, substantial savings in average testing time may be realized for a given overall accuracy level by making decisions sequentially.

ANALYSIS OF ITEM RESPONSE DATA

The analysis of item response data plays quite a different role in criterion-referenced measurement than in classical test development in a norm-referenced context. Traditionally, item analysis has been used to select items that gave the best discrimination among individuals. For purposes of discrimination among people, items should be neither too difficult nor too easy and high correlations of items with the total test score are to be desired. For criterion-referenced measurement, however, such characteristics as item difficulty and discriminating power are generally considered irrelevant. Indeed, when these item statistics are used to select items, they may be viewed as a menace because they can distort the domain representativeness of the test.

Anderson (1972) has persuasively argued that selecting items for their difficulty "...begs the question that a criterion-referenced test is intended to answer" (p. 164). Estimates of a person's level in relationship to a well defined domain is desired rather than an indication that he or she can get more or less items right than other members of some norm group. Selecting items for their discriminating power, also subverts the purpose of a criterion-referenced measure since it redefines the item domain in unspecifiable ways. The more discriminating items may, as suggested by Anderson, result in more of an aptitude measure than an achievement measure. Certainly, the domain is changed which destroys the content validity of the test. As summarized by Anderson, "Manipulating tests to control difficulty level and discriminating power tortures validity" (1972, p. 165).

Item Analysis and Domain Definition

Even if the above position is accepted, item analyses may play a role

in the construction of criterion-referenced measures. Popham and Husek (1969) noted that discrimination indices may be useful as ways if identifying items that need review. An item with a low or even a negative discrimination index would not necessarily be discarded but would be studied for possible flaws. For example, an item might be ambiguous or have a distractor that is better than the keyed answer for more knowledgable examinees. Or, there might be an error in the wording which makes the keyed answer incorrect.

Several people have endorsed Popham and Husek's notion that item analysis results may help identify "bad" items (e.g., Brennan, 1972; Crehan, 1974; Hambleton, Swaminathan, Algina & Coulson, 1975; Wedman, 1973). Advice on what to do with the information, other than study the items for flaws, is more limited, however. If an item that is flagged on the basis of its item statistics is judged to be flawed, in some way one must choose among one of three possible courses of action: (1) retain the item unchanged, (2) modify the item and retain it; (3) discard the item. If either the second or third course of action is taken, a decision about possible redefinitions of the domain is also needed.

Presumably there is not much point in flagging and reviewing items if only the first course of action is to be considered. Hence, we will focus on the latter two possibilities. Ideally, the need to either modify or discard an item should lead to a respecification of the item domain. When completely specified, the domain should lead to the generation of acceptable items only (ignoring errors that may occur between the generation of an item and the final printing of the item on the test booklet). If an item has a distractor that is "better" than the keyed response, there is something wrong with the definition and rules for generating distractors. Correcting this requires some modification of the definition of the domain.

Although the implication that item changes or deletions ideally should be accompanied by domain modifications is clear, it is also clear that domains are seldom completely specified in practice. In the absence of complete specification, flawed items may be useful in pointing to ways in which the specifications need to be tightened. In this latter case, item modifications or deletions should still be accompanied by changes in domain specifications.

This position, while seemingly extreme, is consistent with the position that ideally domain sepcification should be so complete that item generation becomes a completely routinized task. Domain representativeness under these idealized conditions can be pursued by random or stratified random sampling procedures. Discarding an item from a test does not distroy the advantages of such sampling approaches only if the domain is redefined to exclude a sub-category of items typified by the deleted item.

In practice, complete specification is a goal that is pursued but seldom achieved. This is also apt to be the case where items are identified that need to be modified or deleted but the corresponding changes in the domain specifications are unclear.

A final word of caution on the use of item statistics to flag items for review is that item data are a function not only of the item characteristics but of the educational history of the examinees.

Sensitivity to Instruction

The sensitivity of the difficulty of an item to the instructional history of a group of examinees has been a point of concern for criterion-referenced measures. Given the context in which criterion-referenced measurement originated, where the demand was for measures to be used in an individualized-instruction program, the concern about sensitivity to

to instruction seems natural. Another stimulus for the concern, however, comes from a perceived limitation of most available norm-referenced tests for purposes of program evaluation. This perceived limitation is a lack of sensitivity to differences in instructional programs (e.g., Averch, Carroll, Donaldson, Kiesling and Pincus, 1972; Klein, 1971). Furthermore, Kiesling (1975) obtained results which suggested that the criterion-referenced tests developed by the New York State Education Department were more sensitive to amount of instruction received than was a norm-referenced test.

According to Klein and Kosekoff (1973), a criterion-referenced test item "...is considered 'good' if it is sensitive to instruction, that is, if performance on it is related to the degree of instruction obtained" (p.6). A number of indices of the sensitivity of an item to instruction have been suggested (e.g., Cox & Vargus, 1966; Kosekoff & Klein, 1974; Harris & Pearlman, in press; Roudabush, 1973). They have been suggested as indices that might be used for item selection (e.g., Cox & Vargus, 1966; Kosekoff & Klein, 1974; Roudabush, 1973). More recently, however, they have been suggested as useful tools for studying the "teaching-testing complex" but not for purposes of item selection since the latter would destroy the domain representativeness of the items (Harris & Pearlman, in press). We find this latter perspective more in keeping with the goal of rigorous domain definition which enables the use of sampling procedures to construct tests.

Whatever the use to be made of indices of sensitivity to instruction, it must be recognized that such an index does not characterize a property that adheres solely in the item. It is also a function of the effectiveness of instruction. As noted by Harris, Pearlman and Wilcox (in press), there is a confounding of the effects due to characteristics of the items and

the nature of the instructional experiences that cannot be escaped. Thus, an index of item sensitivity to instruction would more accurately be called an index of item sensitivity and/or instructional effectiveness.

The index suggested by Cox and Vargus (1966) is simply the difference between the proportion of people who get an item correct following instruction and the corresponding proportion prior to instruction. If an item is sensitive to instruction, and the instruction is effective, the proportion passing an item following instruction should be high relative to the proportion prior to instruction. The difficulty level of insensitive items, on the other hand, would be expected to remain essentially constant.

Cox and Vargus demonstrated that their index would yield a different ordering of items than a traditional item discrimination index. Hence, if used for the selection of items, the Cox and Vargus index would lead to the selection of a different subset of items than would be selected using a traditional discrimination index. Crehan (1974) reported similar findings. Given the perspective that item analysis is useful for studying the teaching-testing complex and possibly for modifying the definition of the domain but not for item selection, the above results are only of tangential concern. We only would note that the fact that the indices yield different orderings of items indicates that they provide different types of information about the teaching-testing complex.

A limitation of the Cox and Vargus index is that it ignores the type of changes that occur by focusing exclusively on the difficulties at the two points in time. Two items could show equal discrepancies in difficulties prior to and following instruction, yet have quite different patterns of change, both "learning" and "forgetting."

If a dichotomously scored item is given prior to and following instruction, a total of four possible outcomes are possible. These may be

summarized in a two by two contingency table as shown below:

		Following Instruction	
		Fail	Pass
Prior to Instruction	Pass	n_2	n_1
	Fail	n_3	n_4

The n's represent the number of people who had the pattern of pass and fail indicated by the combination of row and column headings for a particular cell. The cells of particular concern for purposes of analyzing sensitivity to instruction are cell 4, which includes people who fail an item prior to instruction but pass it following instruction, and cell 2 which includes people with just the opposite pattern. People included in cell 4 presumably have learned, whereas those in cell 2 presumably have forgotten. Note that such interpretations ignore the possibility of guessing the correct response on one occasion but not the other, or the possibility of responding errors.

For the combination of a sensitive item with effective instruction, n_4 should be large, while n_2 should be small. One of the indices of item sensitivity suggested by Kosekoff and Klein (1974) is simply the difference between the number of "learners," n_4, and the number of "forgetters," n_2, divided by the total number of people ($N = n_1 + n_2 + n_3 + n_4$), i.e., the index is equal to

$$\frac{n_4 - n_2}{N}$$

Kosekoff and Klein also suggested a modification of the above index which provides an adjustment for guessing (also see Roudabush, 1973). The modified index is based on procedure suggested by Marks and Noll (1967)

for use in a different context. The correction procedure requires assumptions that (1) students who know the answer always answer correctly, (2) that those who do not know the answer will guess the correct answer with some probability less than 1.0, (3) that pre-instruction and post-instruction responses are independent for people who do not know the answer, and (4) people who know the answer prior to instruction will also know it following instruction.

Based on the above assumptions, "corrected" frequencies can be estimated for each of the four cells of the contingency table. These "corrected" frequencies are then substituted for their observed counterparts in the sensitivity index shown about to obtain a "corrected" index.[1] The corrected index is based on rather strong assumptions. Also, Kosekoff and Klein found that the "corrected" index provided a rank ordering of items similar to that given by the uncorrected index. Thus, the uncorrected index might be preferred.

Harris and Pearlman (in press) reviewed a variety of potential indices of agreement, association and relative symmetry that have been developed for binary data. Several of these indices were viewed as potentially useful as measures of sensitivity to instruction. The simplest of these is the uncorrected index suggested by Kosekoff and Klein (1974), which Harris and Pearlman refer to as the "index of departure from equal difficulties". Another index that they consider is the "index of departure from relative symmetry" which is the difference between the conditional probability of "learning", $n_4/(n_3 + n_4)$, and the conditional probability of "forgetting,"

[1] The corrected index may be written $(n_4^* - n_2^*)/N$ where the astericks indicate "corrected frequencies which are obtained as follows: $n_2^* = (n_2 + n_3)^2/n_3$ and $n_4^* = (n_2 + n_3)(n_4 - n_2)/n_3$.

$n_2/(n_1 + n_2)$. The latter index may also be modified using the logic laid out by Marks and Noll to obtain a "corrected" conditional probability. As noted by Harris and Pearlman, however, the "corrected" index may be negative and such a result is inconsistent with the probability interpretation.

An alternative to the design requiring testing prior to and following instruction is a "prospective" design (Fleiss, 1973; Harris & Pearlman, in press). In the latter, the test is administered to two groups which are as comparable as possible in all respects except that one group has received instruction in the domain sampled by the test, while the second group has not received such instruction. An obvious problem with this design is that, in practice, the two groups can seldom be assumed to be comparable in all respects except for the presence or absence of the instruction in question (see, Harris & Pearlman). Without this comparability, item sensitivity is confounded not only with instructional effectiveness but also with unknown pre-existing differences between the groups.

Given comparable groups, several indices of sensitivity are potentially informative. Harris and Pearlman suggested three possibilities. These are Pierce's Theta coefficient, the odds ratio described by Fleiss, and Yule's Q, which is a function of the odds ratio. A detailed description of these indices and discussion of some of their properties may be found in Harris and Pearlman. This will not be repeated here, in part, because application of these indices is apt to be quite limited due to the strong assumption of group comparability that is required to interpret them. Also, at this stage there is not sufficient experience with them to give a clear preference for one versus another.

The repeated measures design with testing prior to and following instruction is apt to be the most generally useful approach to studying item sensitivity to instruction. Either the simple "index of departure from equal

difficulties" or the "index of departure from relative symmetry" (Harris & Pearlman, in press) would seem to be potentially useful for such a study. To date, there is limited experience with such studies and the limited results that have been obtained (e.g., Kosekoff & Klein, 1974: CTB/McGraw-Hill, 1973) do not provide much indication of the way indices of item sensitivity might be used to study the "teaching-testing complex".

Item Stability and Item Equivalence

In addition to studies of item sensitivity to instruction, Harris, Pearlman and Wilcox (in press) proposed two other types of studies of item response data as useful to understanding the "teaching-testing complex." These are studies of item stability and studies of item equivalence. Item stability studies, as the name suggests, provide evidence regarding the consistency of item responses on two occasions. Over a brief time interval, a close agreement between the responses on the two occasion is generally desired. An item equivalnce study would investigate the degree to which responses to pairs of items are associated, or are in agreement. For both equivalence and stability studies, Harris and Pearlman recommend the use of the simple proportion of agreement as the index.

Several illustrations of the potential uses of item equivalence and item stability studies were provided by Harris and Pearlman. For example, item equivalence between two items designed to measure the same skill but using different item formats would give an indication of generalizability of the skill over different item formats. A lack of equivalence might suggest that the instruction was too specifically tied to a particular type of presentation.

A variety of indices of association and of agreement that might be used in studies of item stability and/or item equivalence were reviewed by

Harris and Pearlman, but, as previously indicated, they recommend the simple proportion of agreement. The fourth and fifth chapters (Wilcox, in press) of the Harris, Pearlman and Wilcox monograph present new methods of studying stability and of studying equivalence. These methods are developed from the work Lazarsfeld and Kendall (see also Goodman & Kruskal, 1959).

For stability studies, Wilcox presents models for the item response data and associated procedures for estimating "the proportion of examinees who know the answer to the test item" and "the probability that a randomly selected examinee gives an inappropriate response to the test item." An "inappropriate" response is either a wrong response by someone who knows the answer or a correct response by someone who doesn't. For equivalence studies, "the proportion of examinees who know the answer to both test items" or the probability of giving an inappropriate response on either one of them is estimated.

Summary

The analysis of item response data can be of importance to criterion-referenced measures just as it is to norm-referenced ones. The uses of the data that are judged appropriate in the two different contexts are quite different, however. For the criterion-referenced measure, item analyses are apt to be of at least as much relevance to investigations of instructional processes as they are to concerns about item characteristics. Item analyses may provide a means of redefining or sharpening specifications of item domains. In principle, however, item analysis results should not be used for item selection unless such item selection is accompanied by an appropriate redefinition of the domain.

CSE
ARCHIVE COPY

During the past decade there has been some controversy about the procedures for obtaining reliability estimates of criterion-referenced measures. Part of the controversy has stemmed from questions concerning the role of score variability in traditional procedures for estimating reliability versus its role in criterion-referenced measurement. The use, if any, to be made of a criterion score in the estimation of reliability has also been the source of some controversy. Finally, there has been some divergence of opinion about whether reliability should be based on the metric of the measure or on the consistency of the categorizations resulting from use of the measure. Other issues could, of course, be identified, but the other issues can generally be subsumed under, or be viewed as closely related to, one of the above.

Score Variability

As is true with many aspects of criterion-referenced measurement, the roots of the score variability issue can be traced back to discussions in the measurement literature that appeared before criterion-referenced measurement became a major focus of attention. Lindquist, for example, distinguished between tests designed to "...rank the students tested..." and those designed to "...discover specific weaknesses, errors, or gaps in the student's achievement." (1936, p. 20). For the former purpose, score variability would be sought and only discriminating items "...answered by some but not all pupils in the group being tested" (p. 30) would be included in the test. For the diagnostic test, which was described in terms largely in tune with today's descriptions of criterion-referenced measures, however, it was seen as "legitimate to include any items about which information concerning the achievement of pupils is desired" (p. 30) whether or not they were answered correctly or incorrectly by all students in the group to be tested. The

result might be a test with little variability which would appear in an unfavorable light if traditional reliability estimates were used as the standard for judging the test.

Other early discussions related to the issue of score variability for criterion-referenced measures could be noted, but it wasn't until 1969 that Popham and Husek focused attention on score variability specifically within the context of criterion-referenced measurement. Popham and Husek argued that "The issue of score variability is at the core of the difference between norm-referenced and criterion-referenced tests" (1969, p. 3). Variability is essential in order to discriminate among individuals and is something that is sought after in norm-referenced measures, however, for criterion-referenced measures, score variability is seen neither as desirable or undesirable; it is simply irrelevant. "It is, of course, true that one almost always gets variant scores on any psychological test; but that variability is not a necessary condition for a good criterion-referenced test" (Popham & Husek, 1969, p. 3).

Although variability may not be a necessary condition for a good criterion-referenced measure, it is a necessary condition for obtaining high traditional reliability estimates whether they be of the internal consistency, parallel forms, or test-retest variety. The dependence of correlations on variability is well known; reduced variability can generally be expected to yield lower correlations. Thus, reliability estimates based on the correlation of parallel forms or on a test-retest correlation are viewed as suspect for criterion-referenced measures. High correlations would be fine, but are not necessary for good criterion-referenced measures according to this view. In the presence of a low correlation, one would not know whether this was due to a faulty measure or simply to the lack of variability of a perfectly acceptable criterion-referenced measure. Similar comments could be made about traditional measures of internal consistency, since they too depend on variability.

For Popham and Husek the dependence of traditional reliability measures on score variability caused a serious dilemma because internal consistency and stability were viewed as virtues for a criterion-referenced measure, but no clear way of demonstrating the lack of these virtues was seen in the presence of limited-score variability. Their position and acknowledgement of the dilemma with regard to internal consistency was clearly stated.

> "Now it is obvious that a criterion-referenced test should be internally consistent. If we argue that the items are tied to a criterion, then certainly the items should be quite similar in terms of what they are measuring. But, although it is obvious that a criterion-referenced test should be internally consistent, it is not obvious how to assess internal consistency. The classical procedures are not appropriate. This is true because they are dependent on score variability" (Popham & Husek, 1969, p. 5).

While clearly stating the dilemma, Popham and Husek did not provide guidance on how to deal with it other than a half-hearted suggestion that confidence intervals for individual scores might provide a "partial solution". As noted by Simon (1969), however, even the suggested confidence intervals would be dependent upon some type of variability, either within or between persons.

To some, the issue of limited-score variability has been considered to be of little, if any, practical importance. The divergence of opinions was foreshadowed in an earlier debate between Horn (1966, 1968) and Ebel (1968) in the context of tests designed for purposes of assessment rather than prediction. As was later argued for criterion-referenced measures, Horn claimed that the overriding concern for an assessment measure is representativeness of content and the scores on such a measure need not have variance. Ebel, on the other hand, argued that an assessment measure should have variance. He then went on to try to dismiss the issue by noting that "...of course any debate over whether assessment devices <u>should</u> yield variant scores is slightly academic because in practice they almost always <u>do</u> yield variant scores" (Ebel, 1968, p. 71, emphasis in the original). Little harm to the positions would

result from substituting "criterion-referenced measures" for "assessment devices" in the above exchange.

In a more recent exchange, Woodson (1974a, 1974b) suggested that while an item and/or a test may lack variance for a particular sample (e.g., a group of all masters) both items and tests should have variance for "...the population of observations for which a test is designed..." (1974a, p. 63) (e.g., a group consisting of both masters and nonmasters). Millman and Popham (1974), however, responded by saying that "...item and test score variability is an unnecessary requirement" (p. 137) for a criterion-referenced measure.

Both sides of the controversy readily acknowledge that score variability will almost always be found in practice. The crux of the difference lies in the importance attached to the notion that variability is not a necessary consequence of a good criterion-referenced measure. If zero score variability may be expected to occur rarely in practice, then why has it received all of the attention it has? There probably are several answers to this question. In part the debate may be viewed as one of interest "in theory" if not in practice. But, there are more fundamental answers which make the debate an important one. To begin with, a requirement of item and score variability may be seen as a threat to the fundamental principles for the construction of criterion-referenced measures. Clear definition of the content domain and representativeness of the sample of items from that domain are essential to criterion-referenced measurement. If the requirement of item variance resulted in the elimination of selected items from a test, it would destroy the representativeness.

> "The primary purpose of criterion-referenced testing is descriptive wherein the task is to describe the current status of an examinee. The question asked is how well has the examinee performed in relation to some domain of tasks of which the test items are only a sample. This is clearly the use under discussion in the Popham and Husek article..., in Glaser's (1963) classic description of a criterion-referenced test, and in the writings of Ebel (1962) on content standard tests, Osburn (1968) on universe defined tests, and Hively, et. al. (1973) on domain-referenced tests." (Millman & Popham, 1974, p. 137).

Representativeness of the sample for the well-defined domain is the key issue and variability of responses to a sampled item is irrelevant unless used as a guide in the selection of items, in which case, it would be a menace.

Another part of the answer may lie in a belief supported by notions of mastery learning which is in close harmony with much of the thinking about criterion-referenced measures. In mastery learning, the ideal result of instruction in a given content domain would be total mastery by all those instructed and therefore no score variability.

A cynic might note yet another reason. Criterion-referenced measures tend to be quite short compared to the typical norm-referenced measure. A consequence of brevity alone would be generally lower reliability coefficients. Hence, the need to diminish the importance of reliability as estimated by traditional means.

Whatever the reasons for the emphasis on score variability in considering reliability of criterion-referenced measures, it has been important not only to theoretical discussions but to the search for alternate procedures. It has also been a point of concern to the National Education Association. In "Guidelines and Cautions for Considering Criterion-Referenced Testing" published by the NEA, the dependence of traditional reliability estimates on score variability was noted. It was then argued that "Using these techniques causes the tests to discriminate in the same way as do items in standardized tests" (1975, p. 10).

The issue of score variability remains largely unresolved. Attempts to side step it, e.g., Livingston (1972a) whose approach is described below have, not met with general acceptance. It is our view that the issue is a conceptually important one and it should be made clear that variability should not be the guiding principle nor be allowed to distort representativeness. At the same time, however, the use of traditional indices need pose no real danger if not allowed to distort the content coverage. If traditional reliability

indices are high, this is informative. If they are low, they are less informative, especially if the variability of the test scores or a substantial proportion of the items (see Shoemaker, 1969) is small. As with any measure, the importance of various types of reliability evidence will depend upon the uses of and inferences based upon the scores (American Psychological Association, 1974). For some purposes, the consistency of categorizing people as masters or non-masters will be of primary concern (see below). For other purposes, concern will be with the limits of generalization and procedures stemming from the formulation by Cronbach, Gleser, Nanda and Rajaratnam (1972) which should prove most fruitful(see below).

Deviations from a Criterion Score

Livingston (1972a) focused on the notion of a fixed standard or criterion score as the distinguishing characteristic of a criterion-referenced measure. He then used classical concepts of reliability, but with deviations taken around the criterion score rather than the mean, to derive "criterion-referenced reliability" coefficients. Various procedures from classical test theory based on deviations of scores, X, from a mean, μ_X, were simply replaced by counterparts using deviations from the criterion score, C_X. Livingston also showed relationships between his new formulas and their classical counterparts. Thus, for example, the classical definition of reliability is the ratio of the true score variance, σ_T^2, to the observed score variance, σ_X^2, i.e.,

$$\rho_{TX}^2 = \frac{\sigma_T^2}{\sigma_X^2} \qquad (1)$$

The criterion-referenced counterpart derived by Livingston is shown to be equal to:

$$\rho_C^2 = \frac{\sigma_T^2 + (\mu_X - C_X)^2}{\sigma_X^2 + (\mu_X - C_X)^2} \qquad (2)$$

The Livingston formulation was seen as a solution to the problem caused by low score variability. Stanley (1971) noted that "Criterion-referenced tests are meant to be used in situations in which there may be no variation among the

true scores of the examinees; these measures are intended not to discriminate among persons, but to discriminate each person's score from a fixed criterion score. Therefore, a criterion-referenced test can give reliabile information even though its classically defined reliability coefficient equals zero" (p. 435). Stanley then went on to indicate that Livingston's procedure provides the solution to this problem.

The apparent advantage of the Livingston approach can be seen for the special case suggested by Stanley by a simple comparison of equations (1) and (2). When $\sigma_T^2 = 0$, the classical reliability given in (1) is either zero (if $\sigma_X^2 \neq 0$) or is undefined (if $\sigma_X^2 = 0$). The result from equation (2), Livingston's criterion-referenced reliability, will be greater than zero even when $\sigma_T^2 = 0$, as long as $\mu_X \neq C_X$. This apparent advantage of the criterion-referenced reliability msut be weighed against several disadvantages that have been discussed in the literature.

Harris (1972) noted that the fact that Livingston's reliability coefficient is larger (provided $\mu_X \neq C_X$) than the classical reliability coefficient "... does not imply that a more dependable determination of whether or not a true score falls below (or exceeds) a given criterion value" (p. 29). The dependability is a function of the standard error of measurement rather than the reliability and, as Harris showed, the standard error of measurement is the same under the Livingston or the classical formulation. Thus, the larger criterion-referenced reliability may be seen as potentially misleading, or, at least, less informative than the standard error of measurement.[1] Shavelson, Block and Ravitch (1972) showed that the magnitude of the Livingston coefficient can be altered radically as a function of the size of the difference between the mean and the criterion score. As the difference, $\mu_X - C_X$, increases in absolute magnitude so does the magnitude of the Livingston coefficient. Shavelson, et. al.

[1] Harris also made another criticism which is not considered here. See Harris (1972) and Livingston's (1972b) reply.

viewed this dependence on the size of the $\mu_X - C_X$ difference as an important limitation. Livingston (1972c), on the other hand, sees this dependence as quite consistent with the purpose of a criterion-referenced measure. Coupled with the often arbitrary choice of a criterion score (see the section of this paper on setting standards), however, the dependence of the coefficient on its choice would at least seem to demand a better than typical justification for the choice of C_X or the reporting of coefficients for a range of criterion scores.

Possibly the most fundamental objection to the Livingston coefficient is that provided by Hambleton and Novick (1973). They argue that, where the purpose of criterion referenced test is to dichotomize examinees into two categories, e.g., master and nonmaster, then the concern is not with the size of deviation of a person's score from the criterion score but with whether the person is dependably categorized. In the terms of Hambleton and Novick, Livingston employs a "squared error loss function" whereas they would prefer a "threshold loss". In other words, consistency of categorization by say two parellel measures or at two points in time should be evaluated in terms of an all or nothing standard (0 if inconsistent and 1 if consistent), as would be the case with a threshold loss function, rather than by the square of the deviation of a score from the criterion score.

Consistency of Categorization

The framework for much of the more recent work on reliability for criterion-referenced measures was provided by Hambleton and Novick in the article that was just mentioned. They suggested that "...the traditional concepts of reliability and validity could be replaced by a complete decision-theoretic formulation" (1973, p. 168). For situations where a test is used to dichotomize people into two categories, the metric of primary interest for the test is a zero-one score where nonmasters receive a zero and masters a one. Questions of reliability can then be addressed in terms of the zero-one metric rather than in terms of the number-right score or some other metric. Since product-moment

correlations have undesirable properties for zero-one variables, they suggested that "An alternative measure of reliability might simply be the proportion of times the same decision would be made with...two parallel instruments" (Hambleton & Novick, 1973, p. 168).

The basic idea suggested by Hambleton and Novick was developed and extended by Swaminathan, Hambleton and Algina (1974). Specifically, they defined "... reliability of a criterion referenced test as the measure of agreement between the decisions made in repeated test administations" (p. 264). Swaminathan, et. al. considered using the simple proportion of agreement between decisions on the two occasions but rejected this as an index because it "...does not take into account the agreement that could be expected on the basis of chance alone" (p. 264). For example, if a test identified 90% of the people as masters at time 1 and 95% at time 2, simple random selection of the 90% and of the 95% would be expected to identify 85.5% of the people as masters on both occasions and 0.5% of the people as nonmasters on both occasions, for an overall agreement rate of .86.

To adjust for chance agreement, Swaminathan, et. al., proposed that Cohen's (1960) coefficient kappa be used. Kappa is simply the proportion of agreement corrected for chance. It is given by

$$\kappa = \frac{p_o - p_c}{1 - p_c} \quad , \tag{3}$$

where p_c equals the agreement expected by change and p_o is the observed agreement. In the above example, if 89% of the people were identified as masters on both occasions and 1% identified as nonmasters on both occasions, then p_o sould equal .90. As previously indicated p_c would be .86. The resulting Kappa would be .286. The simple proportion of agreement of .90 might yield a rather different impression than the "corrected for chance" proportion of agreement of about .29.

Coefficient kappa is conceptually appealing as an index for a criterion-

referenced measure that is used to categorize people. It can be applied with two or more categories. With more than two categories, the seriousness of the disagreements can be differentially valued by using weighted kappa (Cohen, 1968). Kappa can range from -1 to +1 with a positive one indicating perfect consistency and zero indicating chance agreement. A value of one is possible only when the marginals are equal. Negative values imply that the observed agreement is worse than would be expected on the basis of chance. Sampling variances of kappa and weighted kappa have been derived (Everitt, 1968; Fleiss, Cohen & Everett, 1969) and Light (1973) has proposed a large sample significance test. The relationship between weighted kappa and the intraclass correlation is shown by Fleiss and Cohen (1973).

When the two categorizations are independent in the population, the expected value of kappa is zero. In the non-null case, however, kappa is a biased estimator (Harris & Pearlman, in press). For this and other reasons, Harris and Pearlman prefer the unadjusted proportion of agreement.

Whether kappa or the unadjusted proportion of agreement is used, the approach to reliability suggested by Swaminathan, et. al. (1974) requires dual measurement, either test-retest or parallel forms. For many practical situations this is a severe limitation because it is often feasible to administer only a single test at one point in time. Consequently, there have been several efforts to the extent the Swaminathan, et.al. approach to a single test administration.

Marshall and Haertel (1975; also see Marshall, 1975) proposed a single administration reliability based on the proportion of agreements which they called coefficient beta. Beta is the mean of all possible split-half coefficients of agreement. Thus, it could be obtained by dividing the test into all possible pairs of half-tests. The categorization of examinees on the basis of each half-test in a pair would be compared and the simple proportion of all classifications that agreed would be computed. These proportions would

then be averaged over all split-halfs. To estimate the reliability of the full length test they suggested that beta simply be substituted into an empirically developed formula that was found to approximate the mean of the values obtained using the Spearman-Brown formula. Analogy is apparently the only justification for substitution into the Spearman-Brown formula in the first place and thus the estimate for the full length test is left in doubt.

More recently, two single administration procedures for estimating reliability have been suggested (Huynh, 1976(a); Subkoviak, 1976). Both approaches are based on the idea first put forth by Hambleton and Novick (1973) that the reliability of a criterion-referenced measure is best judged in terms of consistency of the decisions. Thus, they are similar in philosophy to the Swamanathan, et.al. (1974) and Marshall and Haertel suggestions. Huynh follows Swaminathan, et. al. in using kappa while Subkoviak joins Marshall and Haertel in estimating an unadjusted proportion of agreement.

Huynh's approach is derived from a model in which the probability of a given number of items correct for an examinee is assumed to be given by a binomial density function with a parameter equal to the examinee's true score, i.e., the proportion of items in the complete domain that would be answered correctly by the examinee if all items in the domain could be administered. This assumption amounts to assuming that all items in the domain are "exchangable". In other words, it makes no difference which subset of items are selected for an examinee. The true scores are assumed to follow a beta distribution. Together these assumptions yield a "beta-binomial" model which has proved useful in applications of Bayesian statistics (Novick & Jackson, 1974). For large samples, Huynh suggested that the parameters of the beta distribution of true scores be estimated from statistics of the observed test score distribution (see his equations 2 and 3). The joint distribution of the test and a hypothetical parallel form can then be estimated using the beta-binomial model. This estimated joint distribution can then be used to estimate the proportion

of agreement, which in combination with model estimates of the marginal distribution, can be used to estimate kappa.

The model proposed by Huynh is a relatively strong one and requires heavy calculations. The heavy calculations and complexity may discourage many potential users. He did provide an approximation procedure when a moderatly large number of items (say 10 or more) is available, which should extend its utility. But, even 10 items is more than is available on many criterion-referenced tests.

The assumption that the items are exchangeable may be questionable. As suggested by Lord and Novick (1968) the compound binomial model may be better than the simple binomial model. But, "For many purposes, the binomial error model can be viewed as a useful approximation" (Lord & Novick, 1968, p. 524).

The magnitude of kappa is shown to depend on the criterion score used to form categories, on test length, and on score variability (Huynh, 1976(a), pp. 260-261). Kappa tends to be lower for criterion scores near the extremes, to increase with test length, and to increase with test score variability. Since kappa is not unique for a test but depends on the criterion-score that is used to form categories, Millman (1974) recommended that description of the specific situation, including the criterion score used, be included in reporting kappa.

The approach suggested by Subkoviak (1976) has some close similarities to the Huynh appraoch. He too was concerned with consistency of decisions and assumed a binomial model for the observed scores of each examinee based on the examinee's underlying true score. Subkoviak proposed that proportion agreement be used rather than kappa, however. He also suggested some simpler approaches to estimating the true proportion of items correct than the beta distribution approach used by Huynh. The true proportion correct would be estimated either by the observed proportion of items correct for each individual, which is the maximum likelihood estimate, or by a regression estimate based on the observed

sample mean and the sample Kuder-Richardson Formula 21 reliability coefficient. The estimate of the true proportion correct is used in the binomial model to estimate probabilities that a person's score is above or below a given criterion score for the actual test under consideration and for a hypothetical parallel test. Together these estimated probabilities are used to estimate the proportion of agreement.

Subkoviak also has developed procedures based on the compound binomial model in place of the simple binomial model. He compared the estimates using the binomial model and the compound binomial model for a single test administration with estimates based on a dual administration. Both single administration procedures slightly underestimated the dual administration estimates at all cut scores. The compound binomial model was somewhat better than the simple binomial model, but the latter performed quite well and would seem to provide a useful approximation.

The effect of using three different estimates of true proportion correct scores on Subkoviak's proportion of agreement estimates were compared by Noe and Algina (1977). Based on their simulation results, they concluded that regression estimates based on KR-20 provided generally accurate estimates with reasonably small sampling variability.

As was noted above, the magnitude of kappa is dependent on the choice of a criterion score. The same is true of the uncorrected proportion of agreement. Subkoviak obtained higher proportions of agreement for all three estimation procedures that he investigated with criterion scores near the extremes than with intermediate values. In view of the dependence of the coefficients on the choice of criterion score, Huynh (1977) proposed that the maximum value of kappa be estimated over the range of possible criterion scores. The resulting coefficient is called kappamax. He also provides formulas of the standard error of kappamax and approximate formulas for the kappamax of a double-length

test. Although kappamax has the advantage that it would take on only one value for the sample results for a single test rather than different values for each possible criterion score, it does not correspond to any specific decision rule. Consequently, kappamax may be high while the kappa corresponding to a particular decision rule was relatively low. Thus, kappa coupled with a description of the associated decision rule, as was recommended by Millman (1974), has greater practical utility.

Moyer (1977, Moyer & Fishbein, undated) conducted a study comparing reliability results based on traditional procedures (Kuder-Richardson Formula 20) and those based on consistency of classification (kappa). She found that the two approaches yielded very similar information for the tests used in the Michigan Assessment. Similar results were reported by Strasler and Raeth (1977). The tests that had relatively high (low) KR-20's also had relatively high (low) kappas. Moyer and Fishbein (undated) concluded that for practical purposes KR-20 was preferable since it is more familiar and yielded results similar to kappa. Since the two are conceptually quite different, however, this empirical outcome does not provide a persuasive reason for using KR-20 in place of kappa. The choice should be dictated by the purpose (decision consistency or internal consistency) rather than familiarity despite the similarity of results in specific situations.

The Swaminathan, et. al. (1974) paper together with those by Huynh and and Subkoviak provide the means of estimating the degree of decision consistency in either dual or single administration situations. They represent what seems to be a growing consensus that decision consistency is the main reliability issue for criterion-referenced measures. As recently argued by Livingston (1976), however, for some purposes there may be an interest in something besides the reliability of classification. If so, Livingston (1976) argues that the coefficient he previously developed (Livingston, 1972)

provides an alternative where the focus is on the "reliability of measurement" rather than the "reliability of classification".

Additional Alternatives in the Literature

In addition to the proposals made by Livingston (1972a) and those based upon notions of decision consistency, a few other approaches to questions of reliability have been suggested within a criterion-referenced context. Carver (1974) suggested that the reliability of a criterion-referenced measure could be assessed by administering the test to "...two comparable groups" (p. 56) and comparing the percentages of students in the two groups that exceeded a criterion-score. Alternatively, percentages of a single group exceeding a criterion score could be compared (Carver, 1970). These suggestions fail to take into account the extent to which people are consistently classified. The proportion of people in a given category could be the same based on alternate forms without agreement as to who belongs to a given category. Thus, the procedures for getting at decision consistency (e.g., kappa or proportion agreement) are to be preferred.

Carver (1974) also made some suggestions for what he called the "edumetric dimension" of tests. Although distinct from concepts of criterion-referenced measurement, his edumetrics would appear to have relevance in the criterion-referenced context because he was concerned with absolute rather than comparative decisions. He also claimed that his "...edumetric reliability is not dependent on variances of test scores as in psychometric reliability" (p. 514).

According to Carver (1974), reliability "...is conceptually a synonym for consistency" (p. 514). Unlike classical psychometric procedures where there is a concern for consistency of discrimination among persons, Carver suggested that for edumetric purposes the concern is with the consistency of within individual change. He proposed that edumetric reliability be evaluated by comparing the consistency of gains from pretest to posttest on parallel forms administered at both points in time. Alternatively, he suggested that

parallel forms be administered "...under equal treatment conditions..." (p. 514) in order to determine "...the degree to which individual scores are the same on both forms." (p. 514). Haladyna (1975) has argued, however, that change is not the most appropriate focus for achievement tests. Rather, the focus should be on "...whether or not a student has achieved satisfactorily" (p. 603). Thus, consistency of classification decisions is a more important concern for criterion-referenced measurement.

Another point made by Carver (1974) is concerned with the standard error of measurement which was viewed as a psychometric concept and hence not useful for edumetrics. Components of variance will be considered in the following section. Here it should be noted, however, that "...standard errors of measurement are not reliant upon the variability of test scores" (Haladyna, 1975, p. 603). Furthermore, as noted by Cronbach (1975) estimated components of variance are useful not only for comparative decisions but for absolute decisions.

The coefficient of reproducibility was suggested by Cox and Graham (1966) as an alternative to traditional procedures for estimating the reliability of a criterion-referenced measure. The coefficient of reproducibility is appropriate for items that are sequentially scaled such that a person is expected to pass items up to a certain point and fail items beyond that point. In other words, the items should be expected to approximate a Guttman scale (see Torgerson, 1958, Chapter 12). This would not seem to be a reasonable expectation for most criterion-referenced measures. Certainly, it is far removed from the assumption of exchangability of items used by Huynh (1976a). For more on reproducibility and criterion-referenced measurement see Hambleton and Novick (1973, p. 167).

The final approach to be mentioned in this section is the index of efficiency suggested by Harris (1974). The index of efficiency is relevant to notions of reliability but was never intended "...to mean all that reliability

means" (Harris, 1974, p. 106). Since criterion-referenced measures are often used to make dichotomous mastery decisions (i.e., to serve as mastery tests) Harris reasoned that a coefficient that indicated the extent to which a test discriminates between masters and nonmasters would be informative. The index he developed is defined as the ratio of the between-group (i.e., masters vs. nonmasters) sum of squares to the total group sum of squares. The check on the test's ability to discriminate between categories provided by the index of efficiency should be useful. For many purposes, however, the squared deviations used in the index may not be the right metric. As noted by Hambleton, Swaminathan, Algina and Coulson (1975), two non-overlapping distributions of scores for masters and nonmasters could still have substantial within-group variability and therefore an index of efficiency substantially less than one, even though members of the two groups could be discriminated with perfect accuracy.

Generalizability Theory

Generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972) provides a framework for evaluating the dependability of measures that, in our opinion, is relevant not only in a norm-referenced context but in a criterion-referenced one as well. It deserves more attention for addressing reliability questions concerning creterion-referenced measures than it has received to date. In generalizability theory the focus is appropriately on components of variance rather than on coefficients of generalizability. The focus on components of variability may seem contradictory to the rationale of criterion-measures where, as was discussed at length above, score variability is considered irrelevant. Certainly if there is no score variability then all components of variance must also be zero. What use then does generalizability theory have for criterion-referenced measurement? In short, the simple answer is "plenty."

To see the relevance of generalizability theory, it is useful to recall the four types of interpretations based on the results of a measure that are

considered by Cronbach, et. al. Measures may be used (1) to make "absolute decisions," (2) to make a "comparison between two courses of action for individual," (3) to make a "comparison between persons" and (4) to draw a "conclusion about the relation between pairs of variables" (Cronbach, et. al., 1972, pp. 14-15). Only the third type of interpretation falls clearly within the stereotypic mold of a norm-referenced procedure. The first is much more similar to descriptions of criterion-referenced measurement than to norm-referenced measurement.

Cronbach, et. al. described an absolute decision as one

> "Where an individual is to be classified in some way, performance standards may be set which determine how he will be treated. For example, a test is given an applicant for a driver's license, with the predetermined rule that a score of 85% is considered adequate for licensing. In making the decision, this person is considered by himself..." (p. 14, emphasis added).

All that is missing for the devotee of criterion-referenced measurement is the careful description of the item domain and procedures for ensuring representativeness. Adding these would do nothing to change the absolute character of the decision in contrast to a comparative decision. Neither would the addition alter the relevance of the statistical procedures developed by Cronbach, et. al. for application in this context.

The relevance of generalizability theory to criterion-referenced measurement was clearly recognized by Cronbach, et. al. They noted that:

> "Almost all the uses of tests that are now attracting increasing interest demand absolute content-referenced, or criterion-referenced measurement. It is the level of the graduates' performance, and not their standings relative to each other, that one requires to evaluate early childhood education. It is absolute measurement that one uses to prescribe next week's instructional activities in computer-based instruction, and in testing to see if a standard is met. ...The emphasis in generalizability theory is on the universal score itself and not on comparisons of individuals stressed by classical test theory" (p. 384).

Thus, generalizability theory is purported to be of value for criterion-referenced measurement. Indeed there is a natural affinity between the criterion referenced emphasis on domain definition, item sampling for representativeness and making inferences from performance on sampled items about mastery

of a domain and their counterparts with different labels in generalizability theory. Thus, one might refer to the universe of admissible observations, sampling from that universe, and making inferences about a universe score.

The array of estimation procedures encountered in generalizability theory is to great to review here. Basic to all of them, however, is the estimation of components of variance. Components of variance are informative regardless of the amount of score variability. For example, components of variance due to item form, to time of testing or conditions of test administration are all potentially pertinent for a criterion-referenced measure. If item form had a large component of variance associated with it, then one might consider limiting a domain (and thereby the generalization) to a single item form or using item form to stratify for item sampling purposes in order to enhance representativeness.

The relevance of generalizability theory for criterion-referenced measures was recognized in a recent article by Brennan and Kane (1977). The proposal in their article was not simply for an application of generalizability theory, however, but for an adaptation of the theory in a manner parallel to the adaptation of classical reliability estimates that was suggested by Livingston (1972a). Thus, Brennan and Kane argued that the primary interest for a criterion-referenced measure is "...in the difference between each examinee's universe score and the absolute standard defined by the mastery cutting score" (1977, p. 283). Consistent with this interest, Brennan and Kane proposed deviating observed and universe scores from the standard rather than the mean.

The Brennan and Kane approach results in an "index of dependability" that differs from the usual generalizability coefficient as a function of the difference between the cutting score and the mean. Their dependability index is less than the usual generalizability coefficient when the cutting score is near the mean but greater than the usual generalizability coefficent for cutting scores far from the mean.

In generalizability theory, it is the estimate of components of variance rather than generalizability coefficients that are of primary interst. The estimate of the components of variance are not altered by the introduction of a cutting score for purposes of estimating an index of dependability. Thus, the potential utility of generalizability theory does not depend on the acceptance of a cutting score or on a coefficient based on deviations from it.

It is unfortunate that the issue of score variability for criterion-referenced measures has apparently distracted attention from the uses of generalizability theory for such measures. "Estimates of components are particularily illuminating when an instrument is used for absolute measurement" (Cronbach, 1975, p. 602).

Summary

In this section on reliability we have discussed the issue of score variability for a criterion-referenced measure. While concurring with the position that variability is not a necessary requirement for a good criterion-referenced measure, we think that this position has mistakenly been interpreted to imply that certain approaches to assessing the dependability of a measure are inappropriate when the approaches are actually quite appropriate. Generalizability theory is the primary case in point. It has great relevance for criterion-referenced measurement but has received relatively little attention in this context.

We also reviewed a variety of procedures that have been specifically suggested for assessing reliability of criterion-referenced measures. Of these procedures the ones that have built upon the idea of assessing the consistency of decisions (Hambleton & Novick, 1973) appear to us to have the greatest utility. They are most useful in contexts where a measure is used

to categorize people into two or more (but a relatively small number) of groups. Notable in this connection are the papers by Swaminathan, Hambleton and Algina (1974), Huynh (1976a) and by Subkoviak (1976).

On the other hand, when the goal is to determine whether an examinee's true proportion correct for a domain, π, exceeds some specified criterion level, π_0, measures of consistency leave something to be desired. Two measures may yield consistent decisions, yet neither may be very accurate for determining if π is greater than π_0. The accuracy of classification is the primary concern in discussions of problems of determining test length (e.g., Novick & Lewis, 1974). This issue is given more attention in the section of this paper on test length.

ESTIMATION OF DOMAIN SCORES

The definition of domains and the construction of tests to represent those domains are undertaken because there is an interest in measuring a person's ability to perform the tasks defined by the domain. If possible, it would be desirable to know the proportion of all of the items in the domain that a person could answer correctly if they were all administered. But it is usually impossible, or at least inefficient, to administer all possible items in the domain. Consequently, the proportion correct for the entire domain, π, can only be estimated from performance on the sample of items on the test.

The simple proportion correct on the test, p, provides an intuitively obvious estimate of π. The adequacy of p as an estimate of π, as well as the adequacy of other more complicated estimates, depends on the nature of the sample. We have argued elsewhere in this paper that there are great advantages to having a domain specified in such detail that tests can be constructed by simple random sampling of items from the domain. One of the advantages is that, with random sampling, the estimates are interpretable and their statistical properties may be evaluated.

With random sampling, p is not only the intuitively obvious estimate of π, it is also an unbiased maximum-likelihood estimate. Thus, it is directly interpretable and possesses some desirable statistical properties under these conditions.

If only very long tests were of concern, p would probably be considered adequate for any estimation purposes. In practice, however, tests consisting of relatively small numbers of items are most common and under these circumstances, p will have relatively large sampling variability. Thus, better

estimates will often be desired and may be achieved by using collateral information and stronger models.

Information about the group to which an individual belongs may be used to obtain a regression estimate of π. This estimate is a weighted combination of the observed score for an individual and the mean of the group to which the person belongs. The weights are determined by the reliability of the test. The variability of the regression estimates around the domain scores is smaller than that of p. The intuitive appeal of the regression estimate was described by Novick and Jackson (1970) who noted that, for a very short test, the reliability tends to be small which would result in more weight being given to the group mean. As the test gets longer, however, the weight given to the observed score for the individual increases.

Although they acknowledged that the regression estimates are conceptually desirable, Novick and Jackson (1970) noted, that as originally formulated, they require knowledge of the unknown population mean. They then suggested that the problem could be avoided by the use of Bayesian methods of estimation.

Several Bayesian estimation techniques have been formulated (e.g., Lewis, Wang & Novick, 1975; Lord & Stocking, 1976; Novick, Lewis & Jackson, 1973). A detailed review of several of the estimation procedures has also been provided by Hambleton, Swaminathan, Algina & Coulson (1975). The interested reader is referred to these sources for detail on the estimates. Here we only note that a key aspect of the Bayesian techniques is the specification of prior information. If the model holds and the prior information is relatively strong, the Bayesian estimations procedures can result in much more accurate estimates than would be provided by the simple proportion correct, especially when the test is short.

Several empirical Bayes estimation procedures have recently been suggested by Wilcox (in press, c). He used Monte Carlo procedures to

evaluate the empirical Bayes estimates along with the sample mean, $p_.$, for a variety of sets of true scores, π's for either 5 or 10 hypothetical examinees. A compound binomial error modeal was used to generate the responses of the hypothetical examinees for tests of length 5, 10 or 20 items. The empirical Bayes estimates were generally better than the sample mean in terms of a squared error loss function. In some situations the empirical Bayes estimates provided substantial improvements over the simple mean.

While the estimation of π is of general concern, the primary focus in many applications of criterion-referenced measures is on a dichotomous classification, e.g., master and non-master. In such circumstances, the estimate of π may be only of interest as a means to the end of deciding whether π falls above or below a specified criterion level. The sections of this paper on setting standards and test length address the important issues in that context.

If a state model (Meskaukus, 1976) is assumed, then the estimation problem takes on a different character. That is, if it is assumed that π can only take on values of zero or one, a different set of estimation procedures can be used. Macready and Dayton (1977) have recently presented two probalistic models for assessing mastery assuming an all-or-none mastery model. To accomodate response patterns other than all right answers by persons with $\pi = 1$, or all wrong answers by persons with $\pi = 0$, Macready and Dayton assume that non-masters will have some non-zero probability of getting an item right by guessing and masters have some non-zero probability of being "...unable to produce a correct response due to short-term environmental or psychological factors (e.g., he is the victim of 'forgeting')" (1977, p. 100).

Based on the above assumptions, Macready and Dayton presented two models for estimating the proportion of the sample who are masters, the probability of "forgetting," the probability of "guessing" the right answer, and the probability that a person with a given response pattern is a master or a non-master. The models were evaluated for two domains of multiplication items by comparing the observed frequencies of each response pattern with expected frequencies based on the models and their parameter estimates. Although the fit of the model was generally good, a large estimated probability of "forgetting" an item was required for one of the domains. Macready and Dayton concluded that such a large probability of forgetting was "...logically unreasonable..." which suggested "...the possible need for subdividing or otherwise restructuring... [the] domain" (1976, p. 113).

As we have argued in the section of this paper on setting standards, all-or-none models have appeal only for very narrowly defined domains. The domain that Macready and Dayton found in need of possible subdivision was already very limited in scope. It consisted of multiplication items with 2 digit multipliers and either 3 or 4 digit multiplicands which involved "...at least one 'carry' operation for each digit in the multiplier" (Macready & Dayton, 1976, p. 112). If all-or-none models can be reasonably applied only to domains with scopes that are more narrowly defined than the one above, their application would seem to be quite limited indeed.

TEST BIAS

Issues of bias in test use have received considerable attention during the past ten years. It is well known that, on the average, members of some minority groups score lower than majority group members on most widely used aptitude and achievement tests. The existence of differences in average performance is generally agreed to, but the interpretation of the differences is quite controversial. The difference is often attributed to bias in the tests against members of minority groups (e.g., Williams, 1970). Others attribute the differences to unequal experiential backgrounds. In the latter case the tests are presumed to reflect real differences resulting from these inequalities.

For obvious reasons discussion of bias in tests or in their use involve very emotional issues. Testing practices are often viewed as "... obstacles to greater opportunity and status in education and employment, or as more subtle attempts to perpetuate the stereotype of racial and/or ethnic inferiority." (Spencer, 1972, pp. 12, 13).

We will not attempt to deal with the many issues that deserve attention in a detailed discussion of test bias. That is the topic for a separate monograph Our purpose is much more limited. We will only compare the questions of bias that seem most pertinent for criterion-referenced tests to those frequently raised with norm-referenced tests and will consider the putative advantage of criterion-referenced tests in minimizing bias.

Most of the research related to questions of bias in test use has been done in a selection context and there is a rather extensive literature in that context (see, for example, Linn, 1973; Novick & Petersen, 1976; Schmidt, Berner & Hunter, 1973 and references in those sources). Although criterion-referenced tests might be used for selection, that is not their

primary purpose. Thus, much of the literature on bias in test use for selection is only of indirect relevance to criterion-referenced measurement.

Where the purpose of measurement is the assessment of achievement in a well-defined domain, as it is for most criterion-referenced tests, there is seldom a clear external variable that the test is designed to predict and against which the test is most appropriately judged. Thus, the procedures for investigating bias that have been developed in a selection context and which depend on the availability of such an external measure are of little use for criterion-referenced measures. This is the case, however, not only for criterion-referenced measures of achievement but for norm-referenced achievement tests as well.

In the absence of an external variable against which a test is to be judged, questions of bias have generally been addressed by analyses of internal characteristics of the test for minority and majority group persons. The simplest, but least acceptable, approach is to simply compare the mean difficulty for the different groups. Since the groups have different experiential backgrounds and since responses to test items are expected to be influenced by previous experiences, there is not a good a priori basis for assuming that, in the absence of bias, tests should have equal difficulties for different groups. Thus, differences in difficulty do not necessarily imply bias.

Cleary and Hilton (1969) proposed a definition of item bias that has been applied with minor variations in a number of studies. They defined bias in terms of an interaction between items and group membership. According to this conceptualization, items that were unusually difficult (or easy) for members of one group in comparison to their relative difficulty for members of another group would be considered biased. The difficulties of items need not be the same for the groups being compared but, in the absence

of bias, the difference in difficulty would remain constant over items in the subset. This item by group interaction was suggested to be appropriate for considering questions of bias in criterion-referenced tests by Klein and Kosekoff (1973).

There are a number of limitations in using item difficulty and defining bias in terms of the interaction of items with group membership. Part of the problem is due to properties of the difficulty scale (e.g., the difference between difficulties of .5 vs. .7 may not be comparable to differences between .75 and .95). Interpretation also is dependent on the degree of similarity of the discriminating power of the items. For discussion of some of these issues see Angoff and Ford (1973), Hunter (1975), Stanley (1969) and Lord (1977). These limitations apply, however, whether the test is designed as a norm-referenced or criterion-referenced test.

Despite the difficulties, there is evidence that changes in domain definition can affect the magnitude of the difference in the average performance of black students and white students. Medley and Quirk (197) analyzed the interaction between three types of items and group membership. The items were used in an experimental administration of the National Teacher Examination. They were classified as "traditional," "modern" or "black" items. Questions about black authors or composers, for example, were classified as "black" items. They found that the difference in the mean performance of black students and that of white students was smaller on the black and modern items than on the traditional items.

The test investigated by Medley and Quirk was not a criterion-referenced one. No domain was defined in such a way that items could be randomly sampled. Nonetheless, their study raises important questions about the way domains are defined and the implications of alternate definitions for observed group differences. Domains should be defined in reference to an instructional program. If the instructional program includes the study of

few black authors, then the domain, and therefore the test, can be expected to include few items about black authors. Both the instructional program and the domain definition are in need of change in such a case. It is worth bearing in mind, however, that seemingly subtle changes in the domain definition may affect the size of observed group differences.

Another approach to investigating bias of an achievement test was suggested by Green and Roudabush (1976). They define an achievement test as biased if it "... measures at least partly different things for different groups ..." (p. 2). To illustrate this conceptualization, they describe an example where a reading comprehension test is really a measure of comprehension only for persons familiar with the vocabulary. For persons unfamiliar with the vocabulary, the test may be measuring vocabulary and comprehension. Thus, it might be unfairly concluded that members of a minority group are poor in reading comprehension when, for that group, the test was more of a measure of vocabulary than of comprehension.

Green and Roudabush go on to argue that "... criterion-referenced tests should be less biased than ordinary norm-referenced achievement tests ..." (p. 2). Their argument, which follows from earlier work by Green (1973), is based on an expectation that criterion-referenced measures should be less influenced by general background than are norm-referenced tests. This is expected because criterion-referenced tests are (1) more specific in content coverage and (2) are not chosen on the basis of item statistics. The use of item statistics to select items may result in a test that measures broader characteristics that are more influenced by long-term background experiences than by specific short-term achievement objectives. This argument is consistent with Anderson's (1972) claim that item selection on the basis of difficulty and discriminating power is apt to result in a test that is more heavily influenced by aptitude.

Some empirical results are presented by Green and Roudabush in support of their expectation that cirterion referenced tests are apt to be less biased. They found the factor structure of a criterion-referenced test to be more similar for groups of black students and white students than that of a norm-referenced test. They concluded that

> "... standardized norm- referenced achievement test are probably more biased because they measure broader functions than criterion-reference tests. Both are biased to a degree, large or small, that is yet to be assessed since no absolute measure of amount of bias exists. However, the evidence presented here makes it seem reasonable to assume the degree is small in criterion-referenced tests; at least we believe that the amount of bias in the criterion-referenced test we studied is small" (1975 p. 12).

As Green and Roudabush acknowledged, there is no widely accepted definition of bias in an achievement test. Thus, it is hardly a surprise that not all would agree with their conclusion, nor even find their results relevant for the tests they analyzed. Ebel (1975b), for example, argued that there is little reason to expect well-developed achievement tests to be biased whether they are norm-referenced or criterion-referenced. He did not find much support for the belief that "... scores on criterion-referenced tests depend more on specific learnings and less on general background knowledge" (p. 19), yet this is the presumed basis for expecting less bias in criterion-referenced measures.

At this stage, it is premature to conclude that bias is more or less of an issue with criterion-referenced than norm-referenced tests. There is even a lack of concensus on the definition of bias. There are some reasons, however, to hope for fewer difficulties with criterion-referenced measures. Indeed, Ebel's (1975b) suggestions for avoiding bias on any type of achievement test have much in common with perscriptions for developing criterion-referenced tests.

On the other hand, it would seem very naive to assume that many of the underlying issues that provide the stimulus for charges of test bias will

disappear by a switch to criterion-referenced tests. As with other types of educational changes, there is a danger of viewing criterion-referenced measurement as a panacea. It is not. "Lower-socioeconomic groups will score as low on criterion or objective-referenced tests as they do on norm-referenced tests. Basic factors such as malnutrition and lack of motivation toward school and test taking are untouched by change from one type to another" (National Education Association, 1975, p. 12).

MATRIX SAMPLING

When tests are administered for group assessment purposes there can be substantial advantages to using a multiple matrix sampling procedure (see, for example, Lord & Novick, 1968, Ch. 11 ; Shoemaker, 1973). The advantages of simple random sampling of items have been discussed elsewhere in this paper within the context of individual assessment. Random item sampling for group assessment purposes shares these same positive features of interpretability of the estimates and their statistical properties. Improved estimates for the group as a whole can be obtained, however, by randomly dividing the item domain into subsets of items and administering each subset to a different random sample of examinees (Shoemaker, 1975). This procedure whereby different subsets of items are administered to different random samples of examinees is called multiple matrix sampling.

The primary advantage of matrix sampling is that, for a fixed number of items per student, more precise estimates of each parameter (e.g., the group mean or variance) is obtained than would be by administering a single sample of items to all students. That is, the standard error of estimate is smaller using matrix sampling procedures than it is with the traditional procedure of a single sample of items. A consequence of this advantage is that greater breadth and depth can be achieved without an undue amount of testing for any one student (Shoemaker, 1975).

It should be understood that the greater precision of estimates (i.e., greater depth) and the ability to cover more domains (i.e., greater breadth) for a fixed amount of testing time, apply only to the assessment of group characteristics. These advantages may be outweighted where testing is used for both individual and group assessment purposes.

The theoretical advantages of matrix sampling are well known, yet the technique has been used relatively rarely in practice. This probability is due, partially, to the common use of a test for the dual purposes of individual and group assessment. It is also apt to be a consequence of increased administrative difficulty of matrix sampling, which requires the printing of numerous test forms to be administered to different random samples of examinees.

BIBLIOGRAPHY

Anderson, R.C., "How to Construct Achievement Tests to Assess Comprehension," Review of Educational Research, 1972, 42, 145-170.

Baker, E. and Spooner-Smith, L., Evaluation Report, "Making Judgments", Spring, 1976. Center for the Study of Evaluation, University of California, Los Angeles.

Bormuth, J.R., On A Theory of Achievement Test Items. University of Chicago Press, Chicago, Illinois, 1970.

Brennar, R.L., "A Generalized Upper-Lower Item Discrimination Index," Educational and Psychological Measurement, 1972, 32, 289-303.

Cox, R.C., "Item Selection Techniques and Evaluation of Educational Objectives," Journal of Educational Measurement, 1965, 2(2).

Crehan, K., "Item Analysis for Teacher-Made Mastery Tests," Journal of Educational Measurement, 1974, 11, 255-262.

Cronbach, L.S., Rajartnam, N. and Glesser, G.C., "Theory of Generalizability: A Liberalization of Reliability Theory," British Journal of Statistical Psychology, 1963, 16, 137-163.

Cronbach, L., et. al., The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York, Wiley, 1972.

Davis, F. and Diamond, J., "The Preparation of Criterion-Referenced Tests," CSE Monograph Series in Evaluation, No.3, 1975: 116-138. Center for the Study of Evaluation, University of California, Los Angeles.

Derham, C.H., "Criterion-Referenced, Domain-Referenced and Norm-Referenced Measurement: A Parallox View," Educational Technology, 1974, 15, 9-13.

Ebel, R. L., Essentials of Educational Measurement. Englewood Cliffs, New Jersey, Prentice Hall, 1972.

Ebel, R. L., Some limitations of criterion-referenced measurements. Prepared for AERA Symposium. "Criterion-Referenced Measurement: Emerging Issues," Minneapolis, Minnesota, March, 1970.

Follman, J., Review of John R. Bormuth. "On the Theory of Achievement Test Items," Journal of Educational Measurement, 1971, 8(1), 57-59.

Glaser, R., "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, 1963, 18, 519-521.

Glaser, R., and Nitko, A.J., "Measurement in Learning and Instruction," Educational Measurement, 2nd Edition, edited by R.L. Thorndike, Washington, D.C.: American Council on Education, 1971.

Hambleton, R. and Novick, M., "Toward an Integration of Theory and Method for Criterion-Referenced Tests," Journal of Educational Measurement, 1973, 10(3), 159-170.

Harsh, R., "The Forest, Trees, Branches, and Leaves, Revisited -- Norm, Domain, Objectives and Criterion-Referenced Assessments for Educational Assessment and Evaluation," Association for Measurement and Evaluation in Guidance, Monograph 1, 1974.

Hively, W., "Introduction to Domain-Referenced Testing," Educational Technology, 1974, 14, 5-10.

Hively, W., Patterson, H.L. and Page, S.H., "A Universe-Defined System of Arithmetic Achievement Tests," Journal of Educational Measurement, 1968, 5(4), 275-290.

Husek, T.R. and Sirotnik, T., "Item Sampling in Educational Research," Occasional Report, February, 1967, (2).

Klein, S. and Kosecoff, J., "Issues and Procedures in the Development of Criterion-Referenced Tests," ERIC, TM Report 26, September, 1973.

Livingston, S.A., "Criterion-Referenced Application of Classical Test Theory," Journal of Educational Measurement, 1972, 9(1), 13-26.

Macready, G.B. and Mervin, J.C., "Homogeneity Within Item Forms in Domain Referenced Testing," Educational and Psychological Measurement, 1973, 33(2).

Marso, R., "Test Item Arrangement, Testing Time and Performance," Journal of Educational Measurement, 1970, 7(2), 113-118.

Miller, D, "Content, Item Decisions: Orienting Curriculum-Assessment Surveys to Curriculum Management and Modification," Educational Technology, 1974, 11, 29-35.

Millman, J., "Sampling Plan for Domain-Referenced Tests," Educational Technology, 1974, 11, 17-21.

Millman, J., "Criterion-Referenced Measurement," Evaluation in Education: Current Applications. Edited by W. J. Popham, Berkeley, California, McCutchan, 1974.

Millman, J. and Popham, W.J., "Issue of Item and Test Variance for Criterion-Referenced Tests: A Clarification," Journal of Educational Measurement, 1974, 11, 137-138.

Nitko, A., "Problems in the Development of Criterion-Referenced Test: the IPI Pittsburgh Experience," CSE Monograph Series in Evaluation, No.3, 1975: 59-82. Center for the Study of Evaluation, University of California, Los Angeles.

Osburn, H.G., "Item Sampling for Achievement Testing," Educational and Psychological Measurement, 1968, 28, 95-104.

Popham, W.J., Educational Evaluation. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

Popham, W.J., "Selecting Objectives and Generating Test Items for Objective-Based Tests," CSE Monograph Series in Evaluation, No.3, 1975: 13-25, Center for the Study of Evaluation, University of California, Los Angeles.

Popham, W.J and Husek, T.R., "Implications of Criterion-Referenced Measurement," Journal of Educational Measurement, 1969, 6, 1-9.

Sanders, J.R. and Murray, S., "Alternatives for Achievement Testing," Educational Technology, 1976, 17-23.

Saupe, J.L., "Selecting Items to Measure Change," Journal of Educational Measurement, 1966, 3, 223-228.

Sension, D.B. and Rabehl, G., "Test-Item Domains and Instructional Accountability," Educational Technology, 1974, 11, 22-28.

Shavelson, R.J., Block, J.H. and Raviteh, M.M., "Criterion-Referenced Testing: Comments on Reliability," Journal of Educational Measurement, 1972, 9, 133-137.

Skager, R., "Generating Criterion-Referenced Tests from Objective-Based Assessment Systems: Unsolved Problems in Test Development, Assembly, and Interpretation," CSE Monograph Series in Evaluation, No.3, 1975: 47-58. Center for the Study of Evaluation, University of California, Los Angeles.

Shoemaker, P.M., "Toward A Framework for Achievement Testing," Review of Educational Research, 1974, 45, 127-147.

Whitely, E., "Domain Referenced Testing: An Alternative Model for Test Construction," Proceedings of the Annual Convention of the American Psychological Association, 1971, 6, Part 2, 515-516.

Willoughby, L., et. al., A Comparison of Domain-Referenced and Classic Psychometric Test Instruction Methods, 1976. (Ed 131 128).

"END OF DOCUMENT"

"END OF DOCUMENT"